




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

# Capacidad de aprendizaje y razonamiento de ChatGPT en temas de anestesiología

## *ChatGPT's learning and reasoning capacity in anesthesiology*

Gustavo Cruz<sup>a</sup> , Santiago Pedroza<sup>b</sup> , Fredy Ariza<sup>a</sup> <sup>a</sup>Departamento de Anestesiología, Fundación Valle del Lili. Cali, Colombia.<sup>b</sup>Centro de Investigaciones Clínicas, Fundación Valle del Lili. Cali, Colombia.**Correspondencia:** Departamento de Anestesiología, Fundación Valle del Lili, Cra. 98 # 18-49. Cali, Colombia.**E-mail:** [gustavo.cruz@fvl.org.co](mailto:gustavo.cruz@fvl.org.co)

### ¿Qué sabemos acerca de este problema?

El razonamiento médico de ChatGPT ha sido probado en exámenes de parasitología y en preguntas del Examen de Licencia Médica de Estados Unidos. Además, se ha demostrado la viabilidad de ChatGPT como herramienta para la toma de decisiones radiológicas, con el potencial de mejorar el flujo de trabajo clínico. No obstante, el uso de ChatGPT en la toma de decisiones clínicas en casos reales es limitada y su aplicación en el campo de anestesiología es desconocido.

### ¿Qué aporta este estudio de nuevo?

Se 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 mejoría secundaria a un aprendizaje. Este estudio destaca el potencial y posibles usos de los modelos de lenguaje e inteligencia artificial (IA) en anestesiología.

### ¿Cómo citar este artículo?

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

### 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 mejoría secundaria a un aprendizaje.

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

## 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.

## INTRODUCCIÓN

En los últimos años ha surgido un interés creciente por el desarrollo y la aplicación de la inteligencia artificial (IA). Esta nueva tecnología ha transformado la manera como enfrentamos diversas tareas incluyendo el análisis de datos, la automatización industrial y la asistencia virtual, entre otros (1). El avance exponencial en la capacidad de almacenamiento de datos y la creciente digitalización de la información facilita la función principal de la IA; así como analizar grandes cantidades de datos, identificar patrones y emitir una respuesta que normalmente requeriría inteligencia humana para ser completada (2).

La IA ha tenido un impacto significativo en muchos sectores de desarrollo, financiero, humanístico y científico. Aunque su desarrollo en ciencias médicas es prometedor, su aplicación para la atención clínica sigue siendo limitada (3). En el ámbito clínico se observa un aumento progresivo en su utilización, especialmente en la generación de texto. Sin embargo, el incremento de los campos de texto no estructurado, concomitante con la falta de interoperabilidad y comunicación sinérgica entre los sistemas tecnológicos de IA y la infraestructura de

salud, lleva a una carencia sustancial de datos con estructura y legibilidad adecuadas para ser asimilados por los sistemas de IA necesarios para la concepción y desarrollo de algoritmos de aprendizaje profundo (3).

Además, la toma de decisiones en conductas clínicas—que dependen usualmente de múltiples factores—también dificulta la aplicación de esta tecnología. Se ha trabajado con IA en algunas especialidades médicas, aún sin logros trascendentales. Una revisión de 23 estudios sobre la aplicación de IA para la detección del cáncer de mama entre 2010 y 2018 demostró que la mayoría de los estudios fueron retrospectivos y de tamaño reducido, sin la posibilidad de generalizar los resultados (4). Además, una revisión sistemática realizada en 2021 concluyó que la evidencia actual es insuficiente para apoyar la implementación de la IA en la detección temprana del cáncer de mama (5).

ChatGPT (Chat Generative Pre-Trained Transformer) es un modelo de procesamiento del lenguaje natural que se basa en la arquitectura del modelo de lenguaje GPT-3.5 y ha despertado gran interés por su capacidad de generar, comprender e interpretar el lenguaje humano mediante sistemas informáticos (6). Fue desarrollado por

la compañía de investigación en inteligencia artificial Open AI, en San Francisco, California, y fue entrenado con una variante de la arquitectura Transformer, un modelo de aprendizaje profundo con una arquitectura de red neuronal, diseñado para manipular datos secuenciales. Se utilizó un conjunto de datos de 40 GB de texto, resultando con un modelo con 1500 millones de parámetros (7,8). ChatGPT es la última variante de GPT-3 y está diseñado específicamente para interactuar con los usuarios (7). No obstante, se han identificado riesgos asociados con su implementación en el cuidado de la salud, como el sesgo, la privacidad de datos, la información errónea y la falta de referencia adecuada (9).

Huh evaluó el rendimiento de ChatGPT en un examen de parasitología y lo comparó con el de los estudiantes de medicina en Corea. ChatGPT obtuvo una tasa de acierto del 60,8 %, comparado con el promedio de 77 estudiantes de 90,8 % (10). Kung et al., evaluaron el desempeño de ChatGPT en preguntas del Examen de Licencia Médica de Estados Unidos (USMLE) y evidenciaron una capacidad de razonamiento comprensible e ideas clínicas válidas (3). ChatGPT obtuvo un resultado superior al 60 % en el conjunto de datos NBME-Free-Step-1, equi-

valente a una nota aprobatoria para un estudiante de medicina de tercer año (3). Por otro lado, Rao et al. demostraron la viabilidad de ChatGPT como herramienta para la toma de decisiones radiológicas, con el potencial de mejorar el flujo de trabajo clínico (11).

El uso de ChatGPT en la toma de decisiones clínicas en casos reales es limitada y su aplicación en el campo de anestesiología es desconocido. El presente estudio tuvo como 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

**Chat GPT:** Es un sistema de procesamiento de lenguaje natural que utiliza una red neuronal tipo transformador para generar respuestas coherentes y relevantes. El modelo aprende patrones en grandes conjuntos de datos de texto, lo que le permite capturar información contextual y sintáctica en el texto de entrada y así generar respuestas precisas (12). Asimismo, ChatGPT puede ajustarse y adaptarse a diferentes dominios y tareas por medio del entrenamiento con datos específicos. No obstante, el actual

modelo de ChatGPT no tiene la habilidad de buscar en internet al momento de generar respuestas. En cambio, se basa en patrones en sus datos de entrenamiento (11).

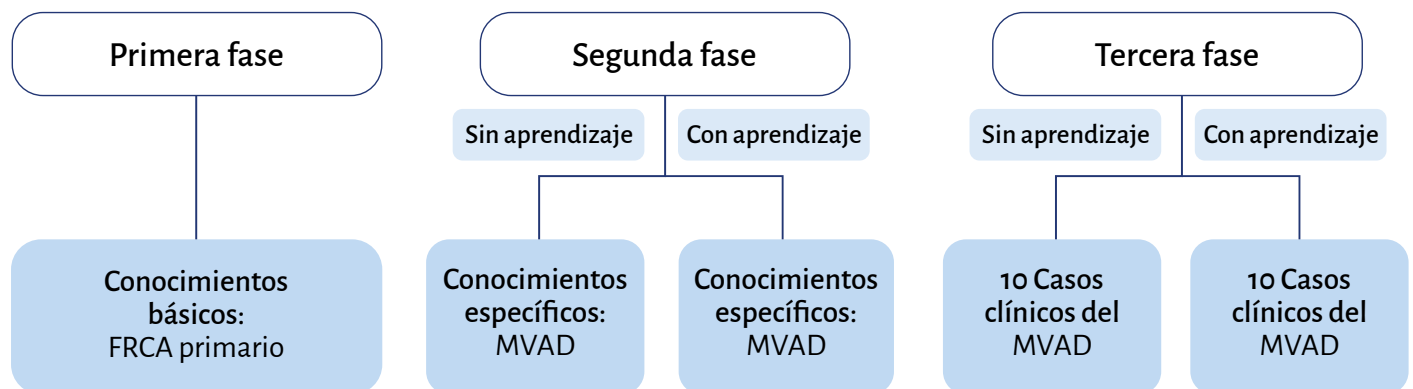
**Fuente de entrada:** Se seleccionaron cuatro fuentes de información. 1) Un banco de preguntas de preparación para el examen primario del Diploma de la Membresía del Colegio Británico Real de Anestesiólogos (FRCA). El FRCA primario es un examen de posgrado, que deben aprobar los anestesiólogos en formación en el Reino Unido antes de postularse para el Entrenamiento Especializado Superior en Anestesiología (13). El examen contiene 45 preguntas de opción múltiple tipo verdadero/falso (OMVF) y 45 preguntas de tipo opción múltiple de única respuesta (OMUR). La tasa de aciertos para aprobar el examen varía cada año, pero normalmente se encuentra entre el 58 % y el 70 % (14). Se extrajeron 840 preguntas de tres libros de bancos de preguntas tipo OMVF para la preparación de este examen (Qbase anesthesia: 1, 2 y 3) (15-17). 2) Examen especializado en el manejo de vía aérea difícil (MVAD). Las preguntas fueron extraídas de Anesthesia HUB (18), una fuente central de anestesiología en la web, en la que se filtraron las preguntas únicamente del MVAD y se obtuvieron 44 preguntas tipo OMUR. 3) Diez reportes de casos clínicos reales del MVAD, extraídos

de PubMed y seleccionados por los investigadores (19-23). 4) Guías de práctica de la Sociedad Americana de Anestesiólogos (ASA) de 2022 para el MVAD (24).

**Método de evaluación:** Se llevó a cabo una evaluación dividida en tres fases (figura 1). En la primera fase se evaluaron los conocimientos del FRCA primario; se realizaron 11 exámenes que contenían 840 preguntas tipo OMVF en las que cada enunciado presentaba 5 opciones de respuesta. En la segunda fase se realizó un examen especializado del MVAD tipo OMUR. Finalmente, en la tercera fase, se evaluaron 10 casos clínicos reales del MVAD, en los que se preguntó acerca de la mejor opción para asegurar la vía aérea y de anestésicos para administrar. La segunda y tercera fase se realizaron antes y después de alimentar a ChatGPT con las guías ASA 2022 del MVAD.

**Codificación/entrada del modelo:** La codificación se organizó en cuatro secciones: 1) Indicaciones (Prompts) de falso o verdadero. Se realizó la entrada textual de la pregunta, con el encabezado: "La siguiente pregunta contiene 5 opciones de respuesta. Responder cada respuesta con falso o verdadero". 2) Preguntas de opción múltiple. Prompts de respuesta única. Se hizo la entrada textual de la pregunta, con el encabezado: "Elegir solo una respuesta (la mejor opción) de la siguiente pregunta". 3) Casos clínicos. Para este tipo de evalua-

**Figura 1.** Fases de la evaluación realizada a ChatGPT.



FRCA: banco de preguntas de preparación para el examen primario del Diploma de la Membresía del Colegio Británico Real de Anestesiólogos, MVAD: manejo de la vía aérea difícil.

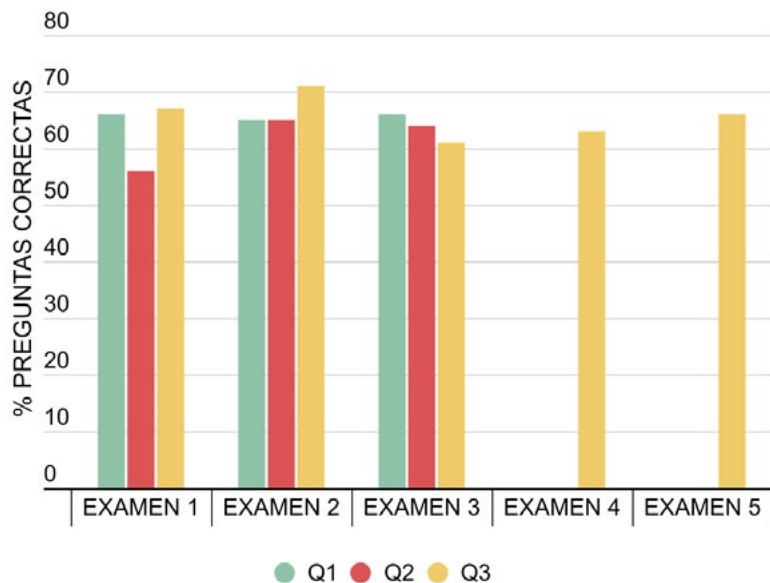
**Fuente:** Autores.

**Tabla 1.** Desempeño de ChatGPT en la evaluación de conocimientos básicos de anestesiología.

Banco de preguntas Q1				
Exámenes	Correctas	Incorrectas	Total	Tasa de acierto (%)
Examen 1 (90)	295	155	450	66
Examen 2 (90)	291	159	450	65
Examen 3 (90)	295	155	450	66
Banco de preguntas Q2				
Exámenes	Correctas	Incorrectas	Total	Tasa de acierto (%)
Examen 1 (90)	256	194	450	56,9
Examen 2 (90)	296	154	450	65,8
Examen 3 (90)	292	158	450	64,9
Banco de preguntas Q3				
Exámenes	Correctas	Incorrectas	Total	Tasa de acierto (%)
Examen 1 (60)	201	99	300	67,0
Examen 2 (60)	215	85	300	71,7
Examen 3 (60)	184	116	300	61,3
Examen 4 (60)	191	109	300	63,7
Examen 5 (60)	200	100	300	66,7

Fuente: Autores.

**Figura 2.** Desempeño de ChatGPT en la evaluación de conocimientos básicos de anestesiología.



Fuente: Autores.

ción se utilizó una “solicitud abierta” o “solicitud sin restricciones”. En esta codificación se eliminaron todas las opciones de respuesta y se agregó una frase interrogativa como encabezado. Después de presentar el caso clínico se formularon dos preguntas de entrada: “En tu opinión, después de leer el

caso clínico, 1. ¿Cuál es la mejor opción (indicar solo una) para asegurar la vía aérea? 2. ¿Cuál sería la mejor opción de anestésicos para asegurar la vía aérea? 4) Alimentación de guías de vía aérea difícil. Se adaptaron las guías ASA 2022 del MVAD en formato escrito (las partes que contenían imágenes

o algoritmos se transcribieron a texto). Posteriormente, se incorporaron las guías por partes (dado que ChatGPT tiene un límite de palabras para transcribir por mensaje), empezando siempre con el enunciado: “aprender esta información y tenerla en cuenta para futuras preguntas”.

**Evaluación de los casos clínicos:** Se creó un método de calificación categorizando las respuestas en tres grupos: concordante, relevante e incorrecta. Las respuestas concordantes correspondieron únicamente al manejo de primera línea o más adecuado, igual al manejo que se eligió en el caso clínico real. Las respuestas relevantes correspondieron a cualquier manejo aceptable. Finalmente, las respuestas incorrectas correspondieron a las opciones de tratamiento erróneas.

Dado que una respuesta concordante también es coherente (pues los manejos de primera línea también entraron en la categoría de un manejo aceptable), todas las respuestas concordantes también fueron clasificadas como coherentes.

Las respuestas de los casos clínicos fueron evaluadas de manera independiente por dos anestesiólogos, quienes desconocían el propósito de la investigación y las respuestas del otro. Se aseguró que ambos médicos desconocieran las evaluaciones mutuas, y se examinó la concordancia en la categorización entre los dos examinadores.

## RESULTADOS

En la evaluación de conocimientos básicos de anestesiología se realizaron 840 preguntas divididas en 11 exámenes. Cada pregunta contenía 5 enunciados de respuesta tipo falso o verdadero, para un total de 4200 respuestas. En esta primera evaluación, Chat GPT obtuvo una tasa de acierto promedio del 65 % (Tabla 1, Figura 2). En el siguiente examen, tipo OMUR, específico en el MVAD, ChatGPT obtuvo una tasa de acierto del 48 % (Tabla 2).

En la última evaluación, las respuestas a las dos preguntas sobre cada caso clínico fueron evaluadas y categorizadas por dos

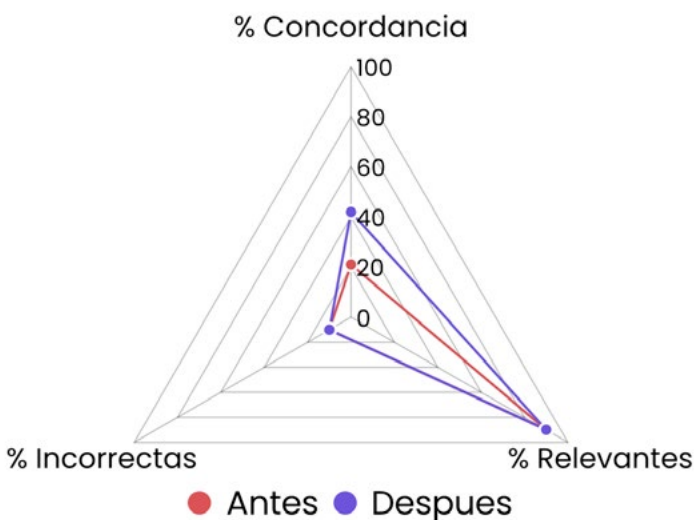
**Tabla 2.** Desempeño de ChatGPT en la evaluación de conocimientos específicos de anestesiología.

Examen específico de vía aérea				
Aprendizaje	Correctas	Incorrectas	Total	Porcentaje correctas
PREA	21	23	44	48
POSTA	26	18	44	59

PREA: previo al aprendizaje, POSTA: posterior al aprendizaje.

Fuente: Autores.

**Figura 3.** Tasas de concordancia, relevancia y falla de ChatGPT en las preguntas de los casos clínicos antes y después del aprendizaje.



Fuente: Autores.

anestesiólogos mediante la inspección del contenido explicativo, con resultado del 100 % de concordancia. ChatGPT formuló respuestas y explicaciones con una concordancia del 20 %, una relevancia del 90 % y una tasa de error del 10 %.

Una vez establecidas la concordancia y la relevancia de las respuestas, se examinó la capacidad de aprendizaje de ChatGPT con el cambio de sus respuestas después de un aprendizaje con las guías ASA 2022 del MVAD (18). Se encontró que en el examen tipo OMUR, ChatGPT mejoró su tasa de acierto al 59 % (Tabla 2). Así mismo, en las respuestas de los casos clínicos, ChatGPT aumentó la concordancia al 40 % (Figura 3).

### DISCUSIÓN

En este estudio se evidencia la capacidad y el potencial de los modelos de lenguaje para la toma de decisiones en diferentes niveles de complejidad en el campo de la anestesiología. ChatGPT mostró 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 una capacidad de mejoría secundario a un aprendizaje.

Estudios previos han evaluado la precisión de ChatGPT para el razonamiento de preguntas médicas. Modelos anteriores (GPT3) registraron un rendimiento del 36,7

% en el Examen de Licencia Médica de los Estados Unidos (USMLE) (25). Pocos meses después, posterior a la actualización de la versión GPT3, Kung et al. demostraron una precisión superior al 50 % en todos los exámenes del USMLE, acercándose al número de aprobación (aproximadamente 60 %) y evidenciando aceptable razonamiento básico y clínico (3). En el presente estudio se informa una precisión similar en la prueba básica de anestesiología tipo OMVF, logrando un puntaje (65 %) dentro del rango superior de la tasa de aciertos necesaria para la aprobación del FRCA primario (58-70 %).

En cuanto al examen específico tipo OMUR sobre el MVAD, el desempeño de ChatGPT no logró alcanzar el umbral de aprobación (superior al 60 %). Una posible explicación para este resultado es el diseño de las preguntas. Al tratarse de una condición anestésica que cuenta con múltiples opciones de tratamiento, incluidas dentro de las opciones de respuesta, solicitar una única respuesta pudo haber disminuido la tasa de aciertos. Esta situación también se presentó en la última prueba, en la que, a pesar de que la tasa de concordancia en las preguntas abiertas de los casos clínicos fue del 20 %, la relevancia de las respuestas — es decir, una respuesta válida alterna a la de primera línea— fue del 90 %.

Resultados similares fueron obtenidos por Yeo et al., quienes evaluaron el desempeño de ChatGPT en preguntas sobre cirrosis y carcinoma hepatocelular. Reportaron el 74 % de las respuestas como completas y comprensibles en las categorías de conocimientos básicos, tratamiento y estilo de vida. Sin embargo, este porcentaje se redujo al 50 % en la categoría de diagnóstico. Tanto el estudio de Yeo et al. como la presente investigación muestran una base sólida de conocimientos en las respuestas de ChatGPT, pero con la limitación en la mayoría de los casos para proporcionar la mejor recomendación individualizada (respuestas válidas, pero no la mejor posible), por lo que se recomienda en la actualidad solo como una herramienta de información complementaria (26).

Se demostró la capacidad de aprendizaje de ChatGPT, mejorando la tasa de aciertos en el examen tipo OMUR y la concordancia en las respuestas de los casos clínicos posterior a la alimentación de las guías de MVAD. Este hallazgo resalta dos puntos importantes. El primero, el inmenso potencial de mejora de ChatGPT u otros modelos de lenguaje en el momento en que se disminuyan las barreras para su alimentación efectiva de información. En segundo lugar, los hallazgos sugieren que las imprecisiones de ChatGPT se atribuyen más a deficiencias informativas que a errores en el procesamiento. Esta premisa se ve respaldada por la alta relevancia observada tanto antes como después del proceso de aprendizaje, así como por el incremento en la concordancia tras dicho proceso. Este último punto también fue descrito por Kung et al., quienes evidenciaron que las respuestas inexactas fueron impulsadas principalmente por información faltante, lo que llevó a una disminución en la comprensión y la indecisión en la IA (3).

Este estudio tiene algunas limitaciones importantes. La cantidad de datos de entrada son relativamente pequeños, en especial en la evaluación de los casos clínicos reales. Esto pudo afectar la profundidad y el alcance de los análisis. Así mismo, un estudio más sólido del modo de falla de la IA (por ejemplo, solicitar una justificación más detallada sobre cada respuesta y evaluar el error en el análisis) podría proporcionar información valiosa sobre la etiología de la inexactitud y la discordancia. Además, algunas limitaciones se derivan de las restricciones propias de ChatGPT. Entre estas, la incapacidad de buscar información nueva/reciente en internet y atribuir información factual a una fuente. Estas limitaciones deben ser consideradas en la evaluación de la toma de decisiones clínicas de esta herramienta.

Es evidente que las herramientas basadas en la tecnología de IA se incorporarán cada vez más en la vida diaria, incluyendo su uso en futuras herramientas especializa-

das en la toma de decisiones clínicas. Este trabajo es fundamental en este proceso, como puerta de entrada a los posibles usos en anestesiología. Así mismo, se destacan las debilidades del estudio y de ChatGPT que pueden ser mejoradas en próximas versiones de esta nueva tecnología.

## CONCLUSIONES

ChatGPT ofrece 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 mejora secundaria a un aprendizaje. Este estudio destaca el potencial y posibles usos de los modelos de lenguaje e IA en anestesiología. Es necesario desarrollar estrategias que permitan minimizar los riesgos de su uso y obtener los mejores beneficios.

## RESPONSABILIDADES ÉTICAS

### Aval de comité de ética

El estudio no fue realizado con animales o con seres humanos, por lo que no se incluye una declaración sobre la aprobación ética en la sección de Métodos.

### Protección de personas y animales

Los autores declaran que para esta investigación no se han realizado experimentos en seres humanos ni en animales. Los autores declaran que los procedimientos seguidos se conformaron a las normas éticas del Comité de Experimentación Humana responsable y de acuerdo con la Asociación Médica Mundial y la Declaración de Helsinki.

### Confidencialidad de los datos

Los autores declaran que han seguido los protocolos de su centro de trabajo sobre la publicación de datos de pacientes.

## Derecho a la privacidad y consentimiento informado

Los autores declaran que en este artículo no aparecen datos de pacientes. Los autores han obtenido el consentimiento informado de los pacientes y/o sujetos referidos en el artículo. Este documento obra en poder del autor de correspondencia.

## RECONOCIMIENTOS

### Contribución de los autores

**GC:** Conceptualización, investigación, escritura, revisión y edición del manuscrito, supervisor.

**SP:** Conceptualización, investigación, escritura, revisión y edición del manuscrito, tablas 1-2 y figuras 1-3.

**FA:** Investigación, escritura, revisión y edición del manuscrito.

### Asistencia para el estudio

Ninguna declarada.

### Apoyo financiero y patrocinio

Ninguno declarado.

### Conflictos de interés

Ninguno declarado.

### Presentaciones

Ninguna declarada.

### Agradecimientos

Ninguno declarado.

## REFERENCIAS

- 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.
- 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>.
- 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>.
- 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>.
- Natalie. Open AI - What is ChatGPT? [internet]. 2023 [citado: 2023 abr 26]. Disponible en: <https://help.openai.com/en/articles/6783457-what-is-chatgpt>
- Royal College of anaesthetists. FRCA Primary MCQ Examination. n.d.
- The Candidate The newsletter for FRCA candidates. n.d.
- Blunt M, Hammond E, McIndoe A. Qbase anaesthesia. Vol. 1, MCQs for the anaesthesia primary. Greenwich Medical Media; 1997.
- Blunt M, Hammond E, McIndoe A. Qbase anaesthesia. Vol. 2, MCQs for the final FRCA. Greenwich Medical Media; 1997.
- Hammond E, McIndoe A. Qbase anaesthesia. Vol. 3, MCQs in medicine for the FRCA. Greenwich Medical Media; 1999.
- Moss D. Anesthesia HUB. EXAMS [internet]. 2013 [citado: 2023 abr 26]. Disponible en: <https://www.anesthesiahub.com/>.
- Hariharasudhan B, Mane R, Cogate 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>.
- 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>.
- 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>.
- 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>.
- 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>.
- 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>.
- 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>.
- 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>.

**Material suplementario.** Concordancia, relevancia y falla de ChatGPT en las preguntas en los casos clínicos sobre vía aérea difícil.

		Previo al aprendizaje			Posterior al aprendizaje		
Casos	Pregunta	Concordante	Relevante	Incorrecto	Concordante	Relevante	Incorrecto
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

**Fuente:** Autores.