Rev Argent Cir. 2026;118(2):e-1961- http://dx.doi.org/10.25132/raac.v118.n2.1961

Chatbots basados en inteligencia artificial: la importancia de saber preguntar

Artificial intelligence-based chatbots: the importance of asking the right questions

M. Elena Peña, Alejandro Gigena, Fernando Iglesia

Servicio de Cirugía de Cabeza y Cuello, División de Cirugía General. Hospital Militar Central. Argentina

Los autores declaran no tener conflictos de interés.
Conflicts of interest None declared.

Correspondencia
Correspondence:
Alejandro M. Zalazar
E- mail:
drzalazaralejandro@gmail.com

RESUMEN

Antecedentes: los *chatbots* basados en inteligencia artificial (CBIA) son una fuente muy utilizada de información médica. La ingeniería de prompts (IP) está orientada a formular y optimizar las preguntas realizadas a los CBIA para mejorar sus respuestas.
Objetivo: evaluar la calidad de los *prompts* y las respuestas generadas por un CBIA en la resolución de casos clínicos de cirugía general, antes y después de una capacitación en IP.
Materiales y métodos: se elaboraron tres casos clínicos ficticios, que debieron ser resueltos por residentes de Cirugía General utilizando ChatGPT-4®. Luego de recibir capacitación en IP, los participantes resolvieron nuevamente los casos. La calidad de los *prompts* se evaluó mediante una escala (5-15 puntos) que consideró completitud, contexto, datos de entrada, formato de salida e instrucción. Las respuestas del *chatbot* se valoraron con una escala (3-15 puntos) que incluyó precisión, completitud y relevancia. Se compararon los resultados antes y después de la capacitación en IP.
Resultados: dieciséis residentes de primero a cuarto año participaron del estudio. La calidad de los *prompts* mejoró significativamente luego de la capacitación en IP, en puntaje total [7,9(1,8) vs. 10,4(2,1), p,<,0,01], completitud, contexto, datos de entrada y formato de salida. También mejoraron las respuestas del *chatbot* en todas las categorías y puntaje total [10,2(2) vs. 11,9(1,8), p,<,0,01].
Conclusión: la capacitación en IP mejoró significativamente la calidad de los *prompts* y las respuestas del CBIA en la resolución de casos clínicos de cirugía general.

Palabras clave: *inteligencia artificial, chatbot, ChatGPT, prompt, Cirugía General.*

ABSTRACT

Background: Artificial intelligence-based chatbots (CBIA) are a widely used source of medical information. Prompt engineering (PE) focuses on designing and optimizing the questions asked of AI/CBs to improve responses.
Objective: The aim of this study was to compare the quality of prompts and the responses provided by an AiCB for clinical case resolution in general surgery, before and following PE training.
Materials and methods: Three fictional clinical cases were developed for residents in general surgery to solve using ChatGPT-4®. After they were trained in PE, the participants solved the cases again. The quality of the prompts was evaluated using a scale (5-15 points) that explored completeness, context, input data, output format, and instructions. The chatbot's answers were assessed using a scale (3-15 points) that included accuracy, completeness, and relevance. The results obtained before and following PE training were compared.
Results: Sixteen postgraduate year 1 to 4 residents participated in the study. The quality of prompts improved significantly following PE training, as assessed by total score [7.9 (1.8) vs. 10.4 (2.1), p< 0.01] for completeness, context, data input, and output format categories. Chatbot's responses also improved across the categories and total score [10.2 (2) vs. 11.9 (1.8), p<0.01].
Conclusion: Training in PE significantly improved the quality of prompts and AiCB's responses for solving general surgery clinical cases.

Keywords: *artificial intelligence, chatbot, ChatGPT, prompt, general surgery.*

Recibido | *Received* 25-12-25
Aceptado | *Accepted* 01-04-26

ID ORCID: M. Elena Peña, 0000-0001-7298-895X; Alejandro Gigena ORCID: 0009-0004-9518-1533; Fernando Iglesia, 0009-0000-5073-2726

Introducción

La inteligencia artificial (IA) ha mostrado un crecimiento notable en los últimos años, especialmente en los grandes modelos de lenguaje (LLM). Estas son herramientas de procesamiento de lenguaje natural (NLP) entrenados mediante aprendizaje automático en grandes cantidades de datos no estructurados. De esta forma, los LLM se utilizan hoy en día para la comprensión de texto, reconocimiento de voz, generación de lenguaje, traducción, etc.[1].

ChatGPT (de Chat-Generative Pre-Trained-Transformer) es un *chatbot* entrenado en el LLM GPT y se destaca por ser capaz de simular conversaciones humanas mediante una interfaz amigable de preguntas y respuestas. Desde su lanzamiento en noviembre de 2022 por OpenAI, ha ganado popularidad en diversas áreas, entre ellas el ámbito médico. Ha demostrado ser útil en la educación de profesionales de la salud[2,3] y como asistente en la toma de decisiones médicas[4-6]. Actualmente, es ampliamente utilizado tanto por profesionales, como pacientes e instituciones de salud[7]. Sin embargo, existen controversias en cuanto a la validez y seguridad de la información que brindan los *chatbot*s basados en inteligencia artificial (CBIA) en el ámbito médico[8,9].

La pregunta, instrucción o palabras que el usuario introduce en el CBIA para obtener una respuesta se denomina prompt. Diversos estudios previos han demostrado que la manera de formular los *prompts* influye significativamente en la calidad de las respuestas que este brinda[10]. La ingeniería de *prompts* (IP) es un campo de investigación dedicado a diseñar y refinar los *prompts* con el objetivo de interactuar más eficazmente con los CBIA y obtener respuestas más adecuadas[1].

Por ser esta un área de conocimiento relativamente nueva y poco conocida por gran parte de los usuarios en el ámbito médico y educativo, surge la pregunta de si el entrenamiento en IP podría influir en la calidad de la información obtenida con un CBIA.

El objetivo de este estudio fue comparar la calidad de los *prompts* generados por residentes y las respuestas brindadas por un CBIA para la resolución de casos clínicos en Cirugía General, antes y después de recibir capacitación en IP. La hipótesis fue que la adquisición de conocimientos en IP mejoraría tanto la calidad de los *prompts* como de las respuestas obtenidas.

Materiales y métodos

Se realizó un estudio de diseño prospectivo, analítico, cuasi-experimental. Como ámbito de trabajo se eligió ChatGPT-4 (última versión al momento del diseño). La población de estudio estuvo conformada por los residentes de Cirugía General y los casos clínicos en Cirugía General diseñados a tal fin. La unidad de análisis fue el conjunto de *prompts* elaborados por los residentes y las respuestas provistas por ChatGPT. El estudio se desarrolló en el período entre junio y julio de 2024.

Los criterios de inclusión de la población de estudio fueron: residentes de todos los años de Cirugía General de un sanatorio de tercer nivel, y casos clínicos sobre patologías frecuentes de cirugía general. No hubo criterios de exclusión.

Procedimientos

Se elaboraron tres casos clínicos ficticios sobre patologías frecuentes en Cirugía General (Anexo 1). El caso clínico 1 se refirió a una paciente joven de sexo femenino que consultó a la Guardia por un cuadro de abdomen agudo quirúrgico. El objetivo de este caso fue evaluar los posibles diagnósticos diferenciales. El caso clínico 2 se trató de un paciente internado en la institución por colecistitis aguda. El objetivo de este caso fue evaluar el tratamiento de la colecistitis de acuerdo con su gravedad y basado en evidencia científica. El caso clínico 3 se trató de un paciente atendido en consultorio ambulatorio por una eventración compleja. El objetivo de este caso fue evaluar la descripción de los pasos de la técnica quirúrgica.

Se convocó a todos los residentes de Cirugía General de un sanatorio de tercer nivel a participar voluntariamente del estudio. Se realizó un único encuentro presencial de 3 horas de duración que se desarrolló en cuatro etapas.

Etapa 1: Evaluación de la experiencia previa de los participantes en el uso de CBIA e IP

Se preguntó a los participantes si habían utilizado alguna vez un CBIA para cualquier actividad, si lo habían utilizado para su actividad profesional médica y si poseían conocimientos sobre IP. Se les envió un formulario de Google para recopilar las respuestas. A continuación, se brindó una breve demostración sobre el uso y los componentes básicos de ChatGPT-4.

Etapa 2: Resolución de casos clínicos

Se entregaron a los participantes los tres casos clínicos ya descriptos. Se les solicitó que resolvieran los casos interactuando con ChatGPT mediante la generación de *prompts*. Se indicó que iniciaran una nueva sesión para cada caso con el fin de evitar retención de memoria por parte del chat y que no modificasen los hiperparámetros del modelo. Se permitió la reformulación del prompt (iteración). Se solicitó que enviaran el prompt generado para cada caso clínico con la respectiva respuesta de ChatGPT mediante un formulario de Google, recopilando un total de tres *prompts* y tres respuestas por cada residente.

Etapa 3: Capacitación en IP

Por último, se brindó capacitación a los residentes sobre IP basándose en la publicación de Meskó[11]. Esto incluyó la definición de IP y su impacto en la obtención de mejores respuestas del CBIA en el ámbito médico. Luego se enumeraron y definieron los elementos básicos de un prompt (instrucción, contexto, datos de entrada y de salida). Finalmente se brindaron estrategias para refinar los *prompts*, por ejemplo, realizar iteraciones, juegos de roles, etcétera.

Etapa 4: Resolución de casos clínicos luego de la capacitación en IP

Se solicitó a los participantes que resolvieran nuevamente los casos clínicos aplicando los conocimientos aprendidos sobre IP. Se solicitó que enviaran el prompt generado para cada caso clínico con la respectiva respuesta de ChatGPT mediante un formulario de Google idéntico al anterior; se recopiló un total de tres *prompts* y tres respuestas adicionales por cada residente.

Análisis de variables

Se evaluó la calidad de los *prompts* elaborados por los residentes y las respuestas provistas por ChatGPT. Luego se compararon los resultados antes y después de recibir la capacitación en IP.

Para el análisis de la calidad de los *prompts* se consideraron cinco categorías: completitud, contexto, datos de entrada, formato de salida e instrucción (Tabla 1). Cada categoría fue evaluada con escalas de Likert y con la sumatoria se calculó el puntaje total de 5 a 15 puntos. La categoría "completitud" recibió el mayor peso relativo en el puntaje final debido a que constituye un requisito fundamental para el razonamiento clínico válido. Esto es particularmente relevante en el ámbito médico, donde la toma de decisiones depende de la disponibilidad de información clínica suficiente y pertinente.

Para el análisis de la calidad de las respuestas se consideraron tres categorías: precisión, completitud y relevancia (Tabla 2). Cada categoría fue evaluada con escalas de Likert y con la sumatoria se calculó el puntaje total de 3 a 15 puntos. En ambas escalas, un puntaje

■ TABLA 1

Descripción de las variables relacionadas con la calidad de los *prompts*

Variable	Definición	Evaluación
Completitud	Inclusión de todos los datos relevantes del caso clínico en el *prompt* (Por ejemplo: "Paciente mujer de 35 años que consulta por dolor abdominal de 6 horas de evolución")	Escala de Likert. Puntaje 1 al 5 (1: Muy incompleto, 2: Incompleto, 3: Moderadamente completo, 4: Completo, 5: Muy completo)
Contexto	Inclusión de datos contextuales del caso clínico en el *prompt* (Por ejemplo: "Soy residente de primer año")	Escala de Likert. Puntaje 1 al 3 (1: No incluye, 2: Incluye parcialmente, 3: Incluye totalmente)
Datos de entrada	Inclusión de información externa sobre la cual basarse en el *prompt* (Por ejemplo: "Responder usando las guías de la Sociedad europea de hernias")	Escala de Likert. Puntaje 1 al 3 (1: No incluye, 2: Incluye con poco detalle,3: Incluye con detalle)
Formato de salida	Inclusión del formato de salida para formular el *prompt* (Por ejemplo: "Responder en 200 palabras")	Escala de Likert. Puntaje 1 al 2 (1: No incluye, 2: Incluye)
Instrucción	Inclusión de instrucciones claras y precisas en el *prompt* (Por ejemplo: "Elaborar una lista de diagnósticos diferenciales")	Escala de Likert. Puntaje 1 al 2 (1: No incluye, 2: Incluye)
Puntaje total prompt	Sumatoria de los puntajes obtenidos en completitud, contexto, datos de salida, datos de entrada e instrucción	Puntaje 5 al 15

■ TABLA 2

Descripción de las variables relacionadas con la calidad de las respuestas

Variable	Definición	Evaluación
Precisión	Veracidad de la información presentada en la respuesta	Escala de Likert. Puntaje 1 al 5 (1: Totalmente incorrecta, 2: Más incorrecta que correcta, 3: Igualmente correcta que incorrecta, 4: Más correcta que incorrecta, 5: Totalmente correcta)
Completitud	Presencia de toda la información necesaria para responder el caso clínico en la respuesta	Escala de Likert. Puntaje 1 al 5 (1: Muy incompleta, 2: Incompleta, 3: Moderadamente completa, 4: Completa, 5: Muy completa)
Relevancia.	Pertinencia de la información presente en la respuesta para responder el caso clínico	Escala de Likert. Puntaje 1 al 5 (1: Totalmente irrelevante, 2: Poco relevante, 3: Moderadamente relevante, 4: Relevante, 5: Totalmente relevante)
Puntaje total respuesta	Sumatoria de los puntajes obtenidos en precisión, completitud y relevancia	Puntaje 3 al 15

superior indica mejor calidad de respuesta o *prompt.*

La evaluación de los *prompts* y de las respuestas fue realizada por dos evaluadores (AG y FI), médicos especialistas en Cirugía General. Ambos fueron previamente entrenados en ingeniería de *prompts* con la misma modalidad que los residentes. Ambos recibieron el mismo set completo de *prompts* y respuestas codificadas, sin información sobre los participantes ni la intervención. Se solicitó que evaluaran los *prompts* y respuestas utilizando las escalas ya mencionadas. Los resultados de ambos evaluadores fueron promediados obteniendo un resultado final.

Para el tratamiento estadístico, las variables cuantitativas se informaron como promedio y desvío estándar para facilitar su interpretación. Las variables cualitativas se expresaron como frecuencias absolutas y relativas. Para el análisis comparativo de los *prompts* y respuestas antes y después de la capacitación se utilizó el test de Wilcoxon para muestras dependientes.

La concordancia interevaluador se evaluó mediante el coeficiente kappa de Cohen ponderado, debido al carácter ordinal de las escalas utilizadas. Se analizaron tanto la puntuación total de los *prompts* y las respuestas como las dimensiones específicas antes y después de la intervención. La magnitud del acuerdo se interpretó según los criterios de Landis y Koch[12].

Se realizó un análisis multivariado mediante regresión ordinal para evaluar la asociación entre la experiencia previa en el uso de CBIA (tanto para actividades generales como para uso médico) y el año de residencia, con la calidad de los *prompts* y de las respuestas generadas por ChatGPT. Las variables dependientes fueron los puntajes totales de calidad de los *prompts* y de las respuestas, evaluados antes y después de la capacitación en ingeniería de *prompts*. El año de residencia se modeló como una variable ordinal, evaluándose su efecto como una tendencia lineal. Se estimaron odds ratios (OR) con sus correspondientes intervalos de confianza y valores de p. Se verificaron los supuestos del modelo, incluyendo la ausencia de colinealidad entre variables independientes mediante el factor de inflación de la varianza (VIF) y el supuesto de regresión paralela mediante el test de Brant. Se consideró la $p < 0,05$ como significativa. El análisis estadístico fue realizado con el programa SPSS v.25® y RStudio®.

Resultados

De los 18 residentes pertenecientes al Servicio de Cirugía General, 16 participaron del estudio. El 75% había utilizado algún CBIA previamente y el 63% lo utilizó para su actividad profesional. Ningún residente poseía conocimientos previos sobre IP (Tabla 3). Todos los residentes resolvieron los tres casos clínicos antes y después de la capacitación en IP, recopilando un total de 96 *prompts* y 96 respuestas.

La calidad de los *prompts* mejoró significativamente luego de la capacitación en IP [Puntaje total antes: 7,9 (1,8); puntaje total después: 10,4 (2,1); $p < 0,01$]. Esta mejoría se observó en las categorías de completitud, contexto, formato de entrada y formato de salida (Tabla 4). En el caso clínico 1 se observó un incremento significativo del puntaje total, completitud, contexto y formato de salida (Anexo 2, Tabla A1). En el caso clínico 2 se observó un incremento significativo del puntaje total, completitud y contexto (Anexo 2, Tabla A3). En el caso clínico 3 se observó un incremento significativo del puntaje total, completitud, contexto y datos de entrada (Anexo 2, Tabla A5)

La calidad de las respuestas mejoró significativamente luego de la capacitación en IP [Puntaje total antes: 10,2 (2); puntaje total después: 11,9 (1,8); $p < 0,01$]. Esta mejoría se observó en las categorías de precisión, completitud y relevancia (Tabla 5). En los casos

■ TABLA 3

Cantidad de residentes según el año de residencia

Variable	n (%)
Participantes	16
Año de residencia	
R1	6 (37%)
R2	3 (19%)
R3	4 (25%)
R4	3 (19%)
Uso previo de CBIA	12 (75%)
Uso previo de CBIA para actividad profesional	10 (63%)
Conocimientos de IP	0

R1: residente de primer año, R2: residente de segundo año, R3: residente de tercer año, R4: residente de cuarto año, CBIA: chatbots basados en inteligencia artificial, IP: ingeniería de prompts.

■ TABLA 4

Variables relacionadas con los prompts antes y después de recibir capacitación en IP

	Antes de IP	Después de IP	p
Completitud	2,2 (1,1)	3,2 (1,3)	<0,01
Contexto	1,3 (0,6)	2,2 (0,7)	<0,01
Datos de entrada	1,3 (0,5)	1,6 (0,8)	<0,01
Formato de salida	1,3 (0,5)	1,6 (0,5)	<0,01
Instrucción	1,8 (0,4)	1,9 (0,3)	0,4
Puntaje total de prompt	7,9 (1,8)	10,4 (2,1)	<0,01

■ TABLA 5

Variables relacionadas con las respuestas antes y después de recibir capacitación en IP

	Antes de IP	Después de IP	p
Precisión	3,6 (0,6)	4,1 (0,7)	<0,01
Completitud	3,6 (0,9)	4,1 (0,8)	<0,01
Relevancia	3,1 (0,9)	3,7 (0,8)	<0,01
Puntaje total respuesta	10,2 (2)	11,9 (1,8)	<0,01

clínicos 1 y 2 se observó un incremento significativo del puntaje total, precisión, completitud y relevancia (Anexo 2, Tablas A2 y A4). En el caso clínico 3 se observó un incremento significativo del puntaje total, precisión y completitud (Anexo 3, Tabla A6).

En el análisis de la concordancia interobservador se encontró que, para el puntaje total de los *prompts*, la concordancia entre evaluadores fue leve antes de la capacitación (κ = 0,04) y aceptable luego de ella (κ = 0,25). Al analizar las dimensiones específicas se observó que la completitud, contexto, datos de entrada e instrucción presentaron niveles de concordancia moderados-sustanciales antes y después de la intervención, mientras que para el formato de salida fue pobre-aceptable (Tabla A7).

En cuanto al puntaje total de las respuestas, la concordancia fue leve antes de la capacitación (κ = 0,04) y pobre luego de esta (κ = −0,11). Al analizar las dimensiones específicas, la concordancia fue pobre para las tres categorías antes y después de la intervención (Tabla A8).

Para el análisis de predictores de calidad se encontró que, en el modelo de regresión ordinal, no se observó asociación entre el año de residencia y el puntaje total de los *prompts* ni de las respuestas generadas por ChatGPT, tanto antes como después de la capacitación en IP.

En cuanto a la experiencia previa en el uso de CBIA para actividades generales, se observó una asociación negativa con el puntaje total de los *prompts* posteriores a la capacitación [OR 0,7 (IC 95%: 0,01-0,57)]. No se observó asociación con el resto de los puntajes para esta variable.

Asimismo, la experiencia previa en el uso de ChatGPT con fines médicos no mostró asociación significativa con el puntaje total de los *prompts* ni de las respuestas en ninguno de los escenarios evaluados (Tablas A9 y A10).

Discusión

En el presente estudio analizamos la calidad de los *prompts* generados por residentes y las respuestas proporcionadas por ChatGPT-4 en la resolución de casos clínicos de Cirugía General, comparando los resultados antes y después de una capacitación en IP. Los hallazgos principales fueron: a) La calidad de los *prompts* mejoró significativamente luego de la capacitación; b) La calidad de las respuestas mejoró significativamente luego de la capacitación; c) El año de residencia y la experiencia previa de los residentes en el uso de CBIA sin capacitación específica en IP no se asociaron a mejor calidad de *prompts* y respuestas.

La IP es un campo de investigación emergente que se basa en el diseño y refinamiento de los *prompts*. El objetivo es interactuar de forma más eficaz con los CBIA y obtener respuestas más adecuadas[1]. Existen múltiples estrategias para lograrlo. Por ejemplo, la técnica zero-shot donde el usuario solicita una tarea al CBIA sin brindarle ejemplos previos, en contraste con la técnica few-shot, donde antes de generar el prompt se brindan ejemplos de cómo debería responder[1,13]. Diversos estudios han puesto en evidencia que el refinamiento de los *prompts* mejora el desempeño de los modelos de lenguaje[14,15]. Sin embargo, en el ámbito médico, especialmente en Cirugía General, este tipo de información es muy escasa y no existen recomendaciones estandarizadas ni guías sobre la interacción óptima con los CBIA[5,11,16]. En nuestro trabajo observamos un impacto positivo de la IP en la calidad de *prompts* creados por los residentes, destacando las categorías de completitud, contexto, datos de entrada, formato de salida e instrucción, las cuales podrían guiar la elaboración de *prompts* médicos en el futuro.

ChatGPT es uno de los CBIA más conocidos y utilizados en la actualidad. Su rendimiento en distintas tareas mejoró significativamente con el avance de los LLM[6,7,17]. Algunos estudios previos analizaron la utilidad de ChatGPT 3.5® en el ámbito de la educación y medicina. En el examen de licencia médica de Estados Unidos (USMLE), por ejemplo, ChatGPT logró resolver más del 50% de las preguntas, equiparando al conocimiento de un estudiante de Medicina de tercer año[18,19]. La versión utilizada en nuestro estudio, ChatGPT-4, ofrece mejoras en precisión y contextualización de respuestas[20-22]. Nuestros resultados sugieren que, aunque esta versión es más precisa, la calidad de sus respuestas depende aún en gran medida de la formulación del prompt, especialmente en entornos médicos donde la precisión y ética de la información son cruciales para evitar consecuencias negativas graves en los pacientes[22].

Es importante destacar que los casos clínicos diseñados abordaron distintos aspectos de la evaluación médica, tales como la elaboración de diagnósticos diferenciales, tratamiento según la evidencia científica y descripción de la técnica quirúrgica. En todos los casos hubo una mejoría de los *prompts* en las categorías de completitud y contexto, lo cual se tradujo en mejores respuestas del *chatbot*. Este hallazgo resalta la relevancia de incorporar información completa y contextualizada en los *prompts*, independientemente del tipo de caso clínico planteado. En contraste, las otras tres categorías mostraron mejoras no significativas tras la capacitación en la mayoría de los escenarios. Particularmente, la categoría instrucción no evidenció cambios en ninguno de los casos. Esto podría explicarse porque los puntajes basales ya eran elevados antes de la capacitación, lo que habría limitado el margen de mejora y reducido la necesidad de enfatizar este aspecto durante el entrenamiento en IP.

Otro aspecto relevante es que, aunque la mayoría de los residentes había utilizado un CBIA en el pasado para fines personales o profesionales, ninguno poseía conocimientos de IP antes de la capacitación. Esto pone de manifiesto la falta de difusión de la IP, pese a

la creciente popularización de estas herramientas. Por otro lado, ni el año de residencia ni la experiencia previa en el uso de ChatGPT, tanto para actividades generales como médicas, se asociaron de forma consistente con una mejor calidad de *prompts* y de respuestas antes y después de la capacitación. Esto sugiere que un nivel formativo mayor y el uso casual de los CBIA no garantizan una interacción más eficaz con estos modelos de lenguaje. En conjunto, tales hallazgos destacan la importancia de incluir conocimientos sobre el uso de CBIA e IP en los programas de formación médica.

Este estudio presenta algunas limitaciones. En primer lugar, se analizó únicamente ChatGPT-4, por lo que los resultados pueden no ser generalizables a otras versiones y modelos de IA. Por otro lado, el uso de casos clínicos ficticios puede limitar la aplicabilidad de estos hallazgos a escenarios de la vida real, con mayor complejidad y variabilidad. Otra limitación fue la baja concordancia interevaluador para las respuestas del CBIA, incluso después de la capacitación. Esto sugiere que su valoración podría estar influida por componentes interpretativos y subjetivos del evaluador. Asimismo, los resultados de la regresión ordinal deben interpretarse con cautela, dado que la amplitud de los intervalos de confianza reflejan una menor precisión en la estimación de los efectos. Finalmente, la muestra de participantes fue pequeña y proveniente de una sola institución, lo cual podría afectar la generalización de los resultados a otros contextos clínicos y formativos.

En conclusión, en este estudio, la capacitación en IP a residentes de Cirugía General mejoró significativamente la calidad de los *prompts* y de las respuestas generadas por ChatGPT. El año de residencia y el uso previo de CBIA sin capacitación específica en IP no se asoció de manera significativa a mejor calidad de *prompts* y respuestas. Todo esto reafirma la importancia de obtener conocimientos en IP para mejorar la eficacia de los *chatbots* de IA en la toma de decisiones médicas. Se recomienda incorporar entrenamiento en IP en los programas de formación médica para maximizar el potencial de estas herramientas, promoviendo su uso seguro y efectivo.

■ ANEXOS

Anexo 1

Caso clínico 1

Consulta a la Guardia una paciente mujer de 35 años por presentar dolor abdominal en fosa ilíaca derecha de 6 horas de evolución (Escala visual analógica: 8/10), sin otros síntomas asociados.

- Antecedentes personales: apendicectomía a los 25 años, colecistectomía a los 30 años.
- Examen físico: tensión arterial 120/80 mm Hg, frecuencia cardíaca 90 lpm, frecuencia respiratoria 18 rpm, temperatura 37 °C. Abdomen blando, depresible, doloroso en fosa ilíaca derecha sin defensa ni dolor a la descompresión. Ruidos hidroaéreos normales. Resto del examen físico sin particularidades.

Exámenes complementarios

- Laboratorio (hemograma, función renal, hepatograma, ionograma): leucocitos 11 000/mm³, resto sin particularidades.
- Ecografía de abdomen: se observa líquido libre en el cuadrante inferior derecho. Resto sin particularidades.

La paciente le pregunta cuál es el diagnóstico y la conducta para seguir.

Caso clínico 2

Usted recibe una interconsulta por un paciente varón de 60 años internado en sala general por dolor abdominal en hipocondrio derecho de 5 días de evolución. Al examen físico se encuentra lúcido, febril (38 °C), taquicárdico (105 lpm), normotenso. Abdomen blando, depresible, doloroso en hipocondrio derecho con defensa y descompresión. En el laboratorio presenta leucocitosis (25 000 glóbulos blancos), sin otra particularidad. En la ecografía de abdomen se observa vesícula biliar con paredes de 5 mm, lito enclavado en bacinete y líquido perivesicular.

Su compañero de guardia le pregunta cuál es el mejor tratamiento para este paciente de acuerdo con la evidencia científica.

Caso clínico 3

Usted se encuentra en el consultorio de cirugía de pared abdominal. Consulta un paciente varón de 58 años por un bulto doloroso en centro abdominal de 2 años de evolución. Presenta episodios frecuentes de dolor que alteran su calidad de vida.

- Antecedentes personales: diabetes, hipertensión arterial, hemicolectomía izquierda por diverticulitis aguda hace 5 años.
- Examen físico: cicatriz mediana infraumbilical. Eventración de línea media con saco de 20 × 20 centímetros.
- Tomografía computarizada (TC) de abdomen y pelvis: eventración centroabdominal con saco de 20 × 20 cm con contenido intestinal y anillo de 12 cm de ancho.

Se diagnostica una eventración compleja. Se le propone al paciente realizar una eventroplastia convencional, pero usted no recuerda del todo los pasos de la técnica quirúrgica.

Anexo 2

■ TABLA A1

Caso clínico 1: variables relacionadas con los *prompts* antes y después de recibir capacitación en IP

	Antes de IP	Después de IP	p
Completitud	2,2 (1,2)	2,9 (1,2)	<0,01
Contexto	1,3 (0,7)	2,2 (0,8)	<0,01
Datos de entrada	1,2 (0,4)	1,4 (0,6)	0,08
Formato de salida	1,3 (0,5)	1,6 (0,5)	0,03
Instrucción	1,8 (0,4)	1,9 (0,3)	0,3
Puntaje total *prompt*	7,75 (2)	10 (1,9)	<0,01

■ TABLA A2

Caso clínico 1: variables relacionadas con las respuestas antes y después de recibir capacitación en IP

	Antes de IP	Después de IP	p
Precisión	3,3 (0,6)	3,9 (0,7)	<0,01
Completitud	3,3 (0,9)	3,8 (0,8)	0,01
Relevancia	2,7 (0,9)	3,5 (0,8)	<0,01
Puntaje total respuesta	9,3 (2,2)	11,3 (1,7)	<0,01

■ TABLA A3

Caso clínico 2: variables relacionadas con los *prompts* antes y después de recibir capacitación en IP

	Antes de IP	Después de IP	p
Completitud	2,2 (1,1)	3,4 (1,4)	<0,01
Contexto	1,3 (0,6)	2,2 (0,7)	<0,01
Datos de entrada	1,4 (0,5)	1,7 (0,9)	0,13
Formato de salida	1,1 (0,3)	1,4 (0,5)	0,06
Instrucción	1,8 (0,4)	1,8 (0,4)	1
Puntaje total prompt	7,8 (2)	10,6 (2,7)	<0,01

■ TABLA A4

Caso clínico 2: variables relacionadas con las respuestas antes y después de recibir capacitación en IP

	Antes de IP	Después de IP	p
Precisión	3,8 (0,7)	4,4 (0,7)	<0,01
Completitud	3,8 (0,8)	4,1 (0,6)	0,2
Relevancia	3,3 (0,8)	4 (0,7)	0,03
Puntaje total respuesta	10,8 (2)	12,4 (1,7)	0,01

■ TABLA A5

Caso clínico 3: variables relacionadas con los prompts antes y después de recibir capacitación en IP

	Antes de IP	Después de IP	p
Completitud	2,3 (1,1)	3,5 (1,3)	<0,01
Contexto	1,3 (0,5)	2,2 (0,8)	<0,01
Datos de entrada	1,4 (0,5)	1,7 (0,8)	0,03
Formato de salida	1,4 (0,5)	1,7 (0,5)	0,09
Instrucción	1,8 (0,4)	1,9 (0,3)	0,7
Puntaje total prompt	8,2 (1,5)	10,9 (1,7)	<0,01

■ TABLA A6

Caso clínico 3: variables relacionadas con las respuestas antes y después de recibir capacitación en IP.

	Antes de IP	Después de IP	p
Precisión	3,7 (0,5)	4 (0,7)	0,03
Completitud	3,7 (0,7)	4,3 (0,9)	0,03
Relevancia	3,3 (0,9)	3,7 (0,8)	0,3
Puntaje total respuesta	10,7 (1,9)	11,9 (1,9)	<0,01

■ TABLA A7

Concordancia interevaluador en el análisis de los *prompts*

	Intervención	Kappa	Interpretación
Completitud	Antes	0,59	Moderada
Completitud	Después	0,67	Sustancial
Contexto	Antes	0,52	Moderada
Contexto	Después	0,59	Moderada
Datos de entrada	Antes	0,7	Sustancial
Datos de entrada	Después	0,77	Sustancial
Formato de salida	Antes	0,23	Aceptable
Formato de salida	Después	-0,12	Pobre
Instrucción	Antes	0,57	Moderada
Instrucción	Después	0,49	Moderada
Puntaje total	Antes	0,04	Leve
Puntaje total	Después	0,25	Aceptable

■ TABLA A8

Concordancia interevaluador en el análisis de las respuestas

	Intervención	Kappa	Interpretación
Precisión	Antes	0,14	Leve
Precisión	Después	0,06	Leve
Completitud	Antes	0,13	Leve
Completitud	Después	-0,07	Pobre
Relevancia	Antes	0,2	Leve
Relevancia	Después	0,04	Leve
Puntaje total	Antes	0,04	Leve
Puntaje total	Después	-0,11	Pobre

■ TABLA A9

Asociación entre el año de residencia, la experiencia previa en el uso de CBIA (general y ámbito médico) y el puntaje total de los *prompts* antes y después de la capacitación en IP

	Intervención	OR (IC 95%)	p
Año de residencia	Antes	1,34 (0,43–4,15)	0,61
Año de residencia	Después	1,95 (0,66–5,81)	0,23
Uso previo CBIA	Antes	1,49 (0,22–10,24)	0,69
Uso previo CBIA	Después	0,07 (0,01–0,57)	0,01
Uso previo CBIA médico	Antes	0,47 (0,08–2,97)	0,42
Uso previo CBIA médico	Después	1,91 (0,31–11,71)	0,46

■ TABLA A10

Asociación entre el año de residencia, la experiencia previa en el uso de CBIA (general y ámbito médico) y el puntaje total de las respuestas antes y después de la capacitación en IP

	Intervención	OR (IC 95%)	p
Año de residencia	Antes	1,76 (0,65–4,73)	0,27
Año de residencia	Después	1,23 (0,43–3,48)	0,7
Uso previo CBIA	Antes	0,59 (0,1–3,61)	0,57
Uso previo CBIA	Después	3,08 (0,41–23,37)	0,28
Uso previo CBIA médico	Antes	1,41 (0,28–6,98)	0,68
Uso previo CBIA médico	Después	0,72 (0,11–4,68)	0,73

■ ENGLISH VERSION

Introduction

Artificial intelligence (AI) has grown remarkably in recent years, especially in the area of large language models (LLMs). A large language model is a natural language processing (NLP) tool that has been trained on vast amounts of data using machine learning. Consequently, LLMs are currently employed for text understanding, speech recognition, language generation, translation, and other related tasks[1].

ChatGPT (Chat-Generative Pre-Trained Transformer) is a chatbot trained on the GPT LLM and is notable for its ability to simulate human conversations through a user-friendly question-and-answer interface. Since its launch in November 2022 by OpenAI, it has gained popularity in various settings, including the medical field. ChatGPT has proven useful for educating healthcare professionals[2,3] and assisting in medical decision-making[4–6]. It is currently widely used by professionals, patients, and healthcare institutions[7]. However, the validity and reliability of the information provided by artificial intelligence-based chatbots (AiBCs) in the medical field remain controversial[8,9].

The question, instruction, or words that users enter into the AiCB to obtain a response are called a prompt. Several previous studies have shown that the way prompts are formulated has a significant impact on the quality of the responses they generate[10]. Prompt engineering (PE) is a field of research dedicated to the design and refinement of prompts to interact more effectively with AiCBs and obtain more appropriate responses[1].

As this is a relatively new field of knowledge with which many users in medical and educational settings are unfamiliar, the question arises whether PE training could influence the quality of information obtained through an AiCB.

The aim of this study was to compare the quality of prompts generated by residents and the responses provided by an AiCB for clinical case resolution in general surgery, before and following PE training. The hypothesis was that gaining knowledge in PE would improve both the quality of prompts and responses obtained.

Materials and methods

We conducted a prospective, analytic and quasi-experimental study. ChatGPT-4 (the most recent version at the time of design) was selected as the platform for this project. The study population included general surgery residents and clinical cases designed for this purpose. The unit of analysis was the set of prompts created by the residents and the responses provided by ChatGPT. The study took place between June and July 2024.

The inclusion criteria for the study population included all the residents in the general surgery program at a tertiary care hospital, and clinical cases involving common general surgery conditions. There were no exclusion criteria.

Procedures

Three fictional clinical cases were developed based on common conditions in general surgery (Appendix 1). Clinical case 1 corresponded to a young female patient who attended the emergency department for acute abdomen requiring surgery. The aim of this case was to evaluate the possible differential diagnoses. Clinical case 2 corresponded to a patient hospitalized for acute cholecystitis. The aim of this case was to assess evidence-based management of cholecystitis according to its severity. Clinical case 3 was a patient attending the outpatient clinic for a complex incisional hernia. The aim of this case was to evaluate the description of the surgical technique steps.

All residents in the general surgery department

at a tertiary care hospital were invited to participate in the study voluntarily. A single 3-hour in-person meeting was held, divided into four stages.

Stage 1: Assessment of participants' prior experience with AiCB and PE

Participants were asked whether they had ever used an AiCB for any activity, whether they had used it in their medical practice, and whether they were aware of PE. A Google form was distributed to collect their responses. Next, the use and basic components of ChatGPT-4 were briefly introduced.

Stage 2: Solving clinical cases.

The three clinical cases described above were distributed to the participants. They were asked to solve the cases by interacting with ChatGPT using prompts. Participants were instructed to start a new session for each case, to prevent the chat system from retaining memory, and to avoid modifying the model's hyperparameters. The residents were allowed to reformulate (iterate) the prompt. They were asked to submit the prompt generated for each clinical case along with ChatGPT's corresponding response via a Google form, resulting in a total of three prompts and three responses per resident.

Step 3: PE training.

Finally, residents were trained on PE based on the publication by Meskó[11]. This included the definition of PE and its impact on obtaining better responses from the AiCB in the medical field. The basic elements of prompts (instructions, context, input data, and output format) were then listed and defined. Finally, strategies for refining the prompts were provided, including iteration and role-playing.

Step 4: Solving clinical cases following PE training

Participants were asked to revisit the clinical cases and apply the knowledge they had gained about PE to resolve them. They were asked to submit the prompt generated for each clinical case, along with ChatGPT's corresponding response, via a Google form identical to the previously submitted, resulting in a total of three prompts and three responses per resident.

Analysis of variables

The quality of the prompts elaborated by the residents and of the responses provided by ChatGPT was evaluated. Then, the results obtained before and following PE training were compared.

Prompt quality was analyzed considering five categories: completeness, context, input data, output format, and instructions (Table 1). Each category was assessed using Likert scales, and the total score — ranging from 5 to 15 points — was calculated by adding up the individual scores. The "completeness" category was assigned the highest relative weight in the final score because it is a fundamental requirement for valid clinical reasoning. This is particularly relevant in the medical field, where decision-making depends on the availability of sufficient and relevant clinical information.

The quality of the responses was analyzed using three categories: accuracy, completeness, and relevance (Table 2). Each category was assessed using

■ TABLE 1

Description of the variables related to prompt quality

Variable	Definition	Evaluation
Completeness	Inclusion of all relevant clinical case data within the prompt (e.g., "35-year-old female patient presenting with abdominal pain that began 6 hours ago...")	Likert scale. Score from 1 to 5 (1: Highly deficient, 2: Incomplete, 3: Partially complete, 4: Substantially complete, 5: Fully comprehensive).
Context	Inclusion of contextual data regarding the clinical case within the prompt (e.g., "I am a postgraduate year-1 resident ").	Likert scale. Score from 1 to 3 (1: Does not include, 2: Partially includes, 3: Fully includes).
Input Data	Inclusion of external information upon which to base the prompt (e.g., "Respond using the guidelines from the European Hernia Society").	Likert scale. Score from 1 to 3 (1: Does not include, 2: Includes with minor detail, 3: Includes with detail).
Output Format	Inclusion of the required output format when formulating the prompt (e.g., "Respond in 200 words").	Likert scale. Score from 1 to 2 (1: Does not include, 2: Includes).
Instruction	Inclusion of clear and precise instructions within the prompt (e.g., "Develop a list of differential diagnoses").	Likert scale. Score from 1 to 2 (1: Does not include, 2: Includes).
Total Prompt Score	Sum of the scores obtained for completeness, context, input data, output format, and instruction.	Score from 5 to 15.

■ TABLE 2

Description of the variables related to response quality

Variable	Definition	Assessment
Accuracy	Accuracy of the information provided in the response	Likert scale Score from 1 to 5 (1: Completely incorrect, 2: Mostly incorrect, 3: Equally correct and incorrect, 4: Mostly correct, 5: Completely correct).
Completeness	Presence of all necessary information within the response to address the clinical case.	Likert scale. Score from 1 to 5 (1: Highly deficient, 2: Incomplete, 3: Partially complete, 4: Substantially complete, 5: Fully comprehensive
Relevance	Relevance of the information provided by the response to address the clinical case	Likert scale Score from 1 to 5 (1: Completely irrelevant, 2: Mostly irrelevant, 3: Partially relevant, 4: Mostly relevant, 5: Completely relevant).
Total response score	Sum of the scores obtained for accuracy, completeness, and relevance.	Score from 3 to 15

Likert scales, and the total score — ranging from 3 to 15 points — was calculated by adding up the individual scores. For both scales, a higher score indicates greater prompt or response quality.

The prompts and responses were assessed by two evaluators (AG and FI), who are general surgeons. Both surgeons had received training in prompt engineering using the same method as that employed with residents. They were administered the same set of prompts and encoded responses, with no information about the participants or the intervention provided. They were asked to assess the prompts and responses using the mentioned scales. The scores calculated by both evaluators were averaged to obtain a final score.

For statistical analysis, quantitative variables were reported as mean and standard deviation to facilitate their interpretation. Qualitative variables were expressed as absolute and relative frequencies. The Wilcoxon test for dependent samples was used to compare prompts and responses before and following the training intervention.

Inter-rater agreement was assessed using Cohen's weighted kappa coefficient, given the ordinal nature of the scales used. The total scores obtained from prompts and responses, as well as the specific dimensions, were analyzed both before and following the intervention. The magnitude of the agreement was interpreted according to the criteria of Landis and Koch[12].

The association between prior experience of AiCB (for both general and medical purposes) and year in the residency program with the quality of prompts and responses generated by ChatGPT was evaluated using ordinal multivariate regression. The dependent variables were the total prompt and response quality scores, assessed before and following prompt engineering training. The year of residency was modeled as an ordinal variable, and its effect was assessed as a linear trend. Odds ratios (OR) were estimated, along with their corresponding confidence intervals and p-values. The model assumptions were verified, including the absence of multicollinearity among independent variables, using the variance inflation factor (VIF), and the assumption of parallel regression using the Brant test. A p-value < 0.05 was considered statistically significant. All the statistical calculations were performed using SPSS v.25® and RStudio® software packages.

Results

Of the 18 residents in the department of general surgery, 16 participated in the study. Seventy-five percent had used an AiCB before, with 63% having used it in their professional activities. None of the residents had any prior knowledge of PE (Table 3). All residents solved the three clinical cases before and following PE training, providing a total of 96 prompts and 96 responses.

Prompts' quality improved significantly following PE training [total score before training: 7.9 (1.8); total score following training: 10.4 (2.1); $p < 0.01$]. This improvement was noted in completeness, context, input data and output format categories (Table 4). In clinical case 1, a significant increase was observed in the total score, completeness, context, and output format categories (Appendix 2, Table A1). In clinical case 2, a significant increase was observed in the total score, completeness and context categories (Appendix 2, Table A3). In clinical case 3, the total score, completeness, context, and input data categories increased significantly (Appendix 2, Table A5).

The quality of responses improved significantly following PE training [total score before training: 10.2 (2); total score following training: 11.9 (1.8); $p < 0.01$]. This improvement was noted in the accuracy, completeness, and relevance categories (Table 5). In clinical cases 1 and 2, the total score, accuracy, completeness, and relevance categories increased significantly (Appendix 2, Tables A2 and A4). In clinical case 3, there was a significant increase in the total score, accuracy, and completeness categories (Appendix 3, Table A6).

The analysis of inter-rater agreement revealed that agreement was low before training ($\kappa = 0.04$) and acceptable following training for the total prompt score ($\kappa = 0.25$). When analyzing the specific dimensions, moderate and substantial levels of agreement were

■ TABLE 3

Number of residents by year of residency

Variable	n (%)
Participants	16
Year of residency	
PGY-1	6 (37%)
PGY-2	3 (19%)
PGY-3	4 (25%)
PGY-4	3 (19%)
Previous use of AiCB	12 (75%)
Previous use of AiCB for professional activities	10 (63%)
PE knowledge	0

PGY-1: post-graduate year 1 resident; PGY-2: post-graduate year 2 resident; PGY-3: post-graduate year 3 resident; PGY-4: post-graduate year 4 resident; AiCB: artificial intelligence-based chatbot; PE: prompt engineering.

■ TABLE 4

Variables related to prompts before and following PE training

	Before PE	Following PE	p-value
Completeness	2.2 (1.1)	3.2 (1.3)	<0.01
Context	1.3 (0.6)	2.2 (0.7)	<0.01
Input data	1.3 (0.5)	1.6 (0.8)	<0.01
Output format	1.3 (0.5)	1.6 (0.5)	<0.01
Instruction	1.8 (0.4)	1.9 (0.3)	0.4
Total prompt score	7.9 (1.8)	10.4 (2.1)	<0.01

■ TABLA 5

Variables related to response before and following PE training

	Before PE	Following PE	p-value
Accuracy	3.6 (0.6)	4.1 (0.7)	<0.01
Completeness	3.6 (0.9)	4.1 (0.8)	<0.01
Relevance	3.1 (0.9)	3.7 (0.8)	<0.01
Total response score	10.2 (2)	11.9 (1.8)	<0.01

found for completeness, context, input data and instructions before and following the intervention, respectively. For the output format, poor and acceptable levels of agreement were found, respectively (Table A7).

The analysis of inter-rater agreement revealed that agreement was low before training (κ = 0.04) and poor following training for the total response score (κ = 0.11). When analyzing the specific dimensions, agreement was poor for all three categories, both before and following the intervention (Table A8).

The analysis of quality predictors revealed no association between the year of residency and the total prompt score or total response score, either before or following PE training in the ordinal regression model.

Regarding previous experience in using AiCB for general activities, there was a negative association with the total prompt score following training [OR 0.7 (95% CI: 0.01–0.57)]. No association occurred with the rest of the scores for this variable.

Similarly, previous experience with ChatGPT for medical purposes was not significantly associated with total prompt or response scores in any scenario (Tables A9 and A10).

Discussion

In this study, we analyzed the quality of prompts generated by residents and the responses provided by ChatGPT-4 for clinical case resolution in general surgery, before and following PE training. The main findings were: a) the prompt quality significantly improved following training; b) response quality significantly improved following the training; c) the year of residency and prior experience of residents in the use of AiCB without specific PE training were not associated with better prompt quality and response quality.

Prompt engineering is an emerging field of research that focuses on designing and refining prompts. The goal is to interact with the AiCB more effectively and obtain more appropriate responses[1]. There are several strategies to achieve this goal. For example, in the zero-shot technique, the user asks the AiCB to perform a task without providing any examples, whereas in the few-shot technique, the user provides examples of how the model should respond before generating the prompt[1,13]. Several studies have shown that refining prompts improves the performance of language models[14,15]. However, in the medical field, particularly in general surgery, such information is scarce, and there are no standardized recommendations or guidelines about optimal interaction with AiCBs[5,11,16]. In our study, we observed that PE produced a positive impact on the quality of prompts created by residents, particularly in the categories completeness, context, input data, output format, and instructions, which could guide the development of medical prompts in the future.

ChatGPT is one of the most widely known and used AiCBs at present. Its performance on various tasks improved significantly with the development of LLMs[6,7,17]. Some previous studies have examined the usefulness of ChatGPT 3.5® in the field of education and medicine. In the United States Medical Licensing Examination (USMLE), for instance, ChatGPT demonstrated capabilities to answer over 50% of the questions, exhibiting a level of knowledge comparable to that of a third-year medical student[18,19]. The version used in our study, ChatGPT-4, offers improved accuracy and contextual relevance of responses[20–22]. Our results suggest that, while this version demonstrates enhanced accuracy, response quality still largely depends on prompt formulation, particularly within medical contexts where the accuracy and ethical integrity of

the information are crucial to avoiding serious adverse outcomes for patients[22].

It is worth noting that the clinical cases presented addressed various aspects of medical evaluation, such as the formulation of differential diagnoses, evidence-based treatments, and descriptions of surgical techniques. The prompts improved in both completeness and context in all cases, resulting in better chatbot responses. This finding underscores the importance of including comprehensive and contextualized information in prompts, regardless of the type of clinical case presented. Conversely, the other three categories showed no significant improvement in most scenarios following training. In particular, the instruction category showed no changes in any of the cases. This could be explained by the fact that baseline scores were already high before training. This may have limited the potential for improvement and reduced the emphasis placed on this aspect during PE training.

Another relevant aspect is that although most residents had already used an AiCB in their personal or professional activities, none had prior knowledge of PE before training. This highlights the PE is not well disseminated, despite the growing popularity of these tools. Conversely, the year of residency and prior experience using ChatGPT, whether for general or medical purposes, were not consistently associated with higher-quality prompts and responses before or following training. This suggests that a higher level of training and casual use of AiCBs do not ensure more effective interaction with these language models. Altogether, these findings underscore the importance of incorporating knowledge on the use of AiCBs and PE into medical training programs.

This study has some limitations. First, we analyzed only ChatGPT-4, so the results cannot be generalized to other versions or other AI models. The use of fictional clinical cases may limit the applicability of these findings to real-world scenarios, which are more complex and variable. Another limitation was the low inter-rater agreement about the responses provided by the AiCB, even following training. This suggests that their assessment might be influenced by their own subjective interpretations. Furthermore, the results of the ordinal regression should be interpreted with caution, as the wide confidence intervals indicate lower accuracy in estimating the effects. Finally, the sample of participants was small and drawn from a single institution, which could limit the generalizability of the results to other clinical and educational settings.

In conclusion, this study found that training general surgery residents in PE significantly improved the quality of prompts and ChatGPT responses. Year of residency and prior AiCB use without specific PE training were not significantly associated with higher-quality prompts or responses. These findings underscore the importance of leveraging PE expertise to enhance the effectiveness of AI-based chatbots in medical decision-making. It is recommended that PE training be incorporated into medical education programs to maximize the potential of these tools and promote their safe and effective use.

■ APPENDICES

Appendix 1

Clinical case 1

A 35-year-old female patient presents to the emergency department with abdominal pain in the right iliac fossa (visual analogue scale: 8/10) that started 6 hours ago, with no other associated symptoms.

- Personal history: appendectomy when she was 25 years old and cholecystectomy at the age of 30.
- Physical examination: blood pressure 12/80 mm Hg, heart rate 90 bpm, respiratory rate 18 bpm, temperature 37 °C. The abdomen is soft, depressible, and tender on palpation in the right iliac fossa, without guarding or rebound tenderness. The bowel sounds are normal. There are no other relevant signs on physical examination.

Ancillary tests:

- Laboratory tests: complete blood count, kidney function, electrolytes: white cell count 11 000/mm^3, with no other abnormal results.
- Abdominal ultrasound: free peritoneal fluid in the right lower quadrant. There were no other abnormalities.

The patient asks about the diagnosis and course of action.

Clinical case 2

You are required for a consultation on a 60-year-old male patient admitted to the general ward with abdominal pain in the right hypochondriac region that has persisted for 5 days. On physical examination, the patient is alert; his body temperature is 38 °C, his blood pressure is normal, and he presents tachycardia (105 bpm). The abdomen is soft, depressible, and tender on palpation in the right hypochondriac region, with guarding and rebound tenderness. The laboratory tests show high white cell count (25,000/mm3) with no other abnormal results. On abdominal ultrasound, the gallbladder walls measure 5 mm, a stone is lodged in

the gallbladder infundibulum, and pericholecystic fluid is visible.

Your on-call colleague asks you about the best treatment for this patient, based on scientific evidence.

Clinical case 3

You are in the abdominal wall surgery clinic. A 58-year-old male patient presents with a painful bulge in the mid-abdominal region that has been present for two years. He complains of frequent episodes of pain that affect his quality of life.

- Personal history: diabetes, hypertension, left hemicolectomy for acute diverticulitis 5 years ago.
- Physical examination: infraumbilical median scar. Midline incisional hernia with a sac measuring 20 × 20 cm.
- Computed tomography scan of the abdomen and pelvis: midline incisional hernia with a 20 × 20 cm sac containing intestinal contents and a 12 cm wide ring.

The diagnosis is complex incisional hernia. The decision is to perform conventional hernia repair, but you don't fully recall the surgical steps.

Appendix 2

■ TABLE A1

Clinical case 1:
prompt-related variables before and following PE training

	Before PE	Following PE	p-value
Completeness	2.2 (1.2)	2.9 (1.2)	<0.01
Context	1.3 (0.7)	2.2 (0.8)	<0.01
Input data	1.2 (0.4)	1.4 (0.6)	0.08
Output format	1.3 (0.5)	1.6 (0.5)	0.03
Instruction	1.8 (0.4)	1.9 (0.3)	0.3
Total prompt score	7.75 (2)	10 (1.9)	<0.01

■ TABLE A2

Clinical case 1:
response-related variables before and following PE training

	Before PE	Following PE	p-value
Accuracy	3.3 (0.6)	3.9 (0.7)	<0.01
Completeness	3.3 (0.9)	3.8 (0.8)	0.01
Relevance	2.7 (0.9)	3.5 (0.8)	<0.01
Total response score	9.3 (2.2)	11.3 (1.7)	<0.01

■ TABLE A3

Clinical case 2:
prompt-related variables before and following PE training

	Before PE	Following PE	p-value
Completeness	2.2 (1.1)	3.4 (1.4)	<0.01
Context	1.3 (0.6)	2.2 (0.7)	<0.01
Input data	1.4 (0.5)	1.7 (0.9)	0.13
Output format	1.1 (0.3)	1.4 (0.5)	0.06
Instruction	1.8 (0.4)	1.8 (0.4)	1
Total prompt score	7.8 (2)	10.6 (2.7)	<0.01

■ TABLE A4

Clinical case 2:
response-related variables before and following PE training

	Before PE	Following PE	p-value
Accuracy	3.8 (0.7)	4.4 (0.7)	<0.01
Completeness	3.8 (0.8)	4.1 (0.6)	0.2
Relevance	3.3 (0.8)	4 (0.7)	0.03
Total response score	10.8 (2)	12.4 (1.7)	0.01

■ TABLE A5

Clinical case 3:
prompt-related variables before and following PE training

	Before PE	Following PE	p-value
Completeness	2.3 (1.1)	3.5 (1.3)	<0.01
Context	1.3 (0.5)	2.2 (0.8)	<0.01
Input data	1.4 (0.5)	1.7 (0.8)	0.03
Output format	1.4 (0.5)	1.7 (0.5)	0.09
Instruction	1.8 (0.4)	1.9 (0.3)	0.7
Total prompt score	8.2 (1.5)	10.9 (1.7)	<0.01

■ TABLE A6

Clinical case 3:
response-related variables before and following PE training

	Before PE	Following PE	p-value
Accuracy	3.7 (0.5)	4 (0.7)	0.03
Completeness	3.7 (0.7)	4.3 (0.9)	0.03
Relevance	3.3 (0.9)	3.7 (0.8)	0.3
Total response score	10.7 (1.9)	11.9 (1.9)	<0.01

■ TABLE A7

Inter-rater agreement in prompt analysis

	Intervention	Kappa	Interpretation
Completeness	Before	0.59	Moderate
Completeness	Following	0.67	Significant
Context	Before	0.52	Moderate
Context	Following	0.59	Moderate
Input data	Before	0.7	Significant
Input data	Following	0.77	Significant
Output format	Before	0.23	Acceptable
Output format	Following	-0.12	Poor
Instruction	Before	0.57	Moderate
Instruction	Following	0.49	Moderate
Total score	Before	0.04	Mild
Total score	Following	0.25	Acceptable

■ TABLE A8

Inter-rater agreement in response analysis

	Intervention	Kappa	Interpretation
Accuracy	Before	0.14	Mild
Accuracy	Following	0.06	Mild
Completeness	Before	0.13	Mild
Completeness	Following	-0.07	Poor
Relevance	Before	0.2	Mild
Relevance	Following	0.04	Mild
Total score	Before	0.04	Mild
Total score	Following	-0.11	Poor

■ TABLE A9

Association between year of residency, prior experience with AiCB (general and medical purposes), and total prompt scores before and following PE training

	Intervention	OR (95% CI)	p-value
Year of residency	Before	1.34 (0.43–4.15)	0,61
Year of residency	Following	1.95 (0.66–5.81)	0,23
Previous use of AiCB	Before	1.49 (0.22–10.24)	0,69
Previous use of AiCB	Following	0.07 (0.01–0.57)	0,01
Previous use of AiCB for medical purposes	Before	0.47 (0.08–2.97)	0,42
Previous use of AiCB for medical purposes	Following	1.91 (0.31–11.71)	0,46

■ TABLE A10

Association between year of residency, prior experience with AiCB (general and medical purposes), and total prompt scores before and following PE training

	Intervention	OR (95% CI)	p-value
Year of residency	Prior	1.76 (0.65–4.73)	0.27
Year of residency	Following	1.23 (0.43–3.48)	0.7
Previous use of AiCB	Prior	0.59 (0.1–3.61)	0.57
Previous use of AiCB	Following	3.08 (0.41–23.37)	0.28
Previous use of AiCB for medical purposes	Prior	1.41 (0.28–6.98)	0.68
Previous use of AiCB for medical purposes	Following	0.72 (0.11–4.68)	0.73

Referencias bibliográficas /References

1. Wang J, Enze S, Sigang Y, Zihao W, Chong M, Haixing D, et al. Prompt Engineering for Healthcare: Methodologies and Applications. ArXiv. 2023; abs/2304.14670. doi:10.48550/arXiv.2304.14670.
2. Goodman RS, Patrinely JR, Stone CA Jr, Zimmerman E, Donald RR, Chang SS, et al. Accuracy and Reliability of Chatbot Responses to Physician Questions. JAMA Netw Open. 2023;6(10):e2336483. doi:10.1001/jamanetworkopen.2023.36483.
3. Sallam M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. Healthcare (Basel). 2023;11(6):887. doi: 10.3390/healthcare11060887.
4. Sorin V, Klang E, Sklair-Levy M, Cohen I, Zippel DB, Balint Lahat N, et al. Large language model (ChatGPT) as a support tool for breast tumor board. NPJ Breast Cancer. 2023;9(1):44. doi:10.1038/s41523-023-00557-8.
5. Kuşcu O, Pamuk AE, SütaySüslü N, Hosal S. Is ChatGPT accurate and reliable in answering questions regarding head and neck cancer? Front Oncol. 2023;13:1256459. doi:10.3389/fonc.2023.1256459.
6. Hirosawa T, Harada Y, Yokose M, Sakamoto T, Kawamura R, Shimizu T. Diagnostic Accuracy of Differential-Diagnosis Lists Generated by Generative Pretrained Transformer 3 Chatbot for Clinical Vignettes with Common Chief Complaints: A Pilot Study. Int J Environ Res Public Health. 2023;20(4):3378. doi:10.3390/ijerph20043378.
7. Cascella M, Semeraro F, Montomoli J, Bellini V, Piazza O, Bignami E. The Breakthrough of Large Language Models Release for Medical Applications: 1-Year Timeline and Perspectives. J Med Syst. 2024;48(1):22. doi:10.1007/s10916-024-02045-3.
8. Birkun AA, Gautam A. Large Language Model (LLM)-Powered Chatbots Fail to Generate Guideline-Consistent Content on Resuscitation and May Provide Potentially Harmful Advice. Prehosp Disaster Med. 2023;38(6):757–63. doi: https://doi.org/10.1017/S1049023X23006568.
9. Zúñiga Salazar G, Zúñiga D, Vindel CL, Yoong AM, Hincapie S, Zúñiga AB, et al. Efficacy of AI Chats to Determine an Emergency: A Comparison Between OpenAI's ChatGPT, Google Bard, and Microsoft Bing AI Chat. Cureus. 2023;15(9):e45473. doi: https://doi.org/10.7759/cureus.45473.
10. Liu M, Okuhara T, Chang X, Shirabe R, Nishiie Y, Okada H, et al. Performance of ChatGPT Across Different Versions in Medical Licensing Examinations Worldwide: Systematic Review and Meta-Analysis. J Med Internet Res. 2024;26:e60807. doi:10.2196/60807.
11. Meskó B. Prompt Engineering as an Important Emerging Skill for Medical Professionals: Tutorial. J Med Internet Res. 2023;25:e50638. doi:10.2196/50638.
12. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159-74.
13. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172-80. doi:10.1038/s41586-023-06291-2.
14. Wang L, Chen X, Deng X, Wen H, You M, Liu W, et al. Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. NPJ Digit Med.

2024;7(1):41. doi:10.1038/s41746-024-01029-4.
15. White J, Fu Q, Hays S, Sandborn M, Olea C, Gilbert H, et al. A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. ArXiv. 2023:2302.11382v1. https://doi.org/10.48550/arXiv.2302.11382
16. Sorin V, Klang E, Sklair-Levy M, Cohen I, Zippel DB, Balint Lahat N, et al. Large language model (ChatGPT) as a support tool for breast tumor board. NPJ Breast Cancer. 2023;9(1):44. doi:10.1038/s41523-023-00557-8.
17. Open AI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, et al. GPT-4 Technical Report ArXiv. 2023: 2303.08774v6. doi.org/10.48550/arXiv.2303.08774.
18. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. 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 Med Educ. 2023;9:e45312. doi:10.2196/45312.
19. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2(2):e0000198. doi:10.1371/journal.pdig.0000198.
20. Rojas M, Rojas M, Burgess V, Toro-Pérez J, Salehi S. Exploring the Performance of ChatGPT Versions 3.5, 4, and 4 With Vision in the Chilean Medical Licensing Examination: Observational Study. JMIR Med Educ. 2024;10:e55048. doi: 10.2196/55048.
21. Shieh A, Tran B, He G, Kumar M, Freed JA, Majety P. Assessing ChatGPT 4.0's test performance and clinical diagnostic accuracy on USMLE STEP 2 CK and clinical case reports. Sci Rep. 2024;14(1):9330. doi: 10.1038/s41598-024-58760-x.
22. Guillen-Grima F, Guillen-Aguinaga S, Guillen-Aguinaga L, Alas-Brun R, Onambele L, Ortega W, et al. Evaluating the Efficacy of ChatGPT in Navigating the Spanish Medical Residency Entrance Examination (MIR): Promising Horizons for AI in Clinical Medicine. Clin Pract. 2023;13(6):1460-87. doi: 10.3390/clinpract13060130.