RESEARCH ARTICLE

VOL. 44 | NO. 4 | **SPECIAL ISSUE 2023** | PP 152 - 164

Revista Mexicana de Ingeniería Biomédica

https://doi.org/10.17488/RMIB.44.4.11

E-LOCATION ID: 1390

Evaluación de la Calidad de los Agentes Conversacionales para la Creación de Instrumentos de Evaluación en Medición de Señales Bioeléctricas

Evaluation of the Quality of Conversational Agents for the Creation of Evaluation Instruments in Bioelectric Signals Measurement

Alberto Isaac Pérez-Sanpablo[1,2], *Marcela D. Rodriguez-Urrea*[3], *María del Carmen Arquer-Ruíz*[2]
Adrian Octavio Ramirez-Morales[2], *Alicia Meneses-Peñaloza*[1]

[1] Instituto Nacional de Rehabilitación Luis Guillermo Ibarra Ibarra - México

[2] Universidad La Salle - México

[3] Universidad Autónoma de Baja California - México

RESUMEN

El objetivo de esta investigación es evaluar la calidad de agentes conversacionales basados en Modelos de Lenguaje Grandes, para la evaluación de aplicación de conocimiento en Ingeniería Biomédica. Se desarrolló un instrumento de evaluación sobre seis temas de medición de señales bioeléctricas elaborado por un agente humano y los agentes conversacionales Chat-GPT y Bard. Se evaluó la calidad del instrumento en términos de nivel de pensamiento, validez, relevancia, claridad, dificultad y capacidad de discriminación, mediante índice kappa (k) del acuerdo de dos expertos y análisis Rasch de resultados de treinta y ocho estudiantes. Tras eliminar siete preguntas de los agentes conversacionales por problemas de validez y originalidad se integró un instrumento de seis preguntas. Las preguntas fueron válidas y relevantes, claras (>0.95, k=1.0), con dificultad baja a alta (0.61-0.87, k=0.83), índice de discriminación adecuado (0.11-0.47), a nivel de pensamiento de análisis (k=0.22). El promedio de los estudiantes fue de 7.24±2.40. Este es el primer análisis crítico de la calidad de los agentes conversacionales a un nivel de pensamiento superior al de comprensión. Los agentes conversacionales presentaron limitaciones en términos de validez, originalidad, dificultad y discriminación en comparación con el experto humano lo que resalta la necesidad aún de su supervisión.

PALABRAS CLAVE: Bard, Chat-GPT, evaluación educativa, ingeniería biomédica, inteligencia artificial

ABSTRACT

This research aims to evaluate the quality of conversational agents based on Large Language Models for evaluating the application of knowledge in Biomedical Engineering. An evaluation instrument was developed on six topics for measuring bioelectrical signals prepared by a human agent and the conversational agents Chat-GPT and Bard. The quality of the instrument was evaluated in terms of level of thinking, validity, relevance, clarity, difficulty, and discrimination capacity, using the kappa (k) index of the agreement of two experts and Rasch analysis of results from thirty-eight students. After eliminating seven questions from the conversational agents due to validity and originality problems, a 6-question instrument was integrated. The questions were valid and relevant, clear (>0.95, k=1.0), with low to high difficulty (0.61-0.87, k=0.83), adequate discrimination index (0.11-0.47), at the analysis level of thinking (k =0.22). The average score of the students was 7.24±2.40. This is the first critical analysis of the quality of conversational agents at a level of thinking higher than comprehension. The conversational agents presented limitations in terms of validity, originality, difficulty, and discrimination compared to the human expert, which highlights the need for their supervision.

KEYWORDS: artificial intelligence, Bard, biomedical engineering, Chat-GPT, educational measurement

Autor de correspondencia

DESTINATARIO: **Alberto Isaac Pérez-Sanpablo**

INSTITUCIÓN: **Institution Instituto Nacional de Rehabilitación Luis Guillermo Ibarra Ibarra**

DOMICILIO: **Calz México-Xochimilco 289, Coapa, Col. Arenal de Guadalupe, Tlalpan, 14389 Ciudad de México, México**

CORREO ELECTRÓNICO: **albperez@inr.gob.mx**

Recibido:

31 oct 2023

Aceptado:

18 dic 2023

INTRODUCCIÓN

Los avances recientes en inteligencia artificial (IA) y Modelos de Lenguaje Grandes (LLMs, por sus siglas en inglés) [1], están permitiendo desarrollar agentes conversacionales capaces de asistir al humano en tareas que requieren de conocimiento especializado. Ejemplos de estos agentes son los modelos de lenguaje ChatGPT y Bard [2], reemplazado por Gemini en diciembre del 2023, que están siendo altamente adoptados para apoyar actividades de educación, investigación y la práctica de las Ciencias Médicas [3].

Dado que los modelos de lenguaje de estos agentes fueron generados mediante un entrenamiento con una gran cantidad de datos, son capaces de generar respuestas coherentes, pero no necesariamente con un alto grado de precisión para todos los dominios de aplicación. Por lo anterior, es importante evaluar su desempeño para asistir en tareas académicas en el ámbito de la Ingeniería Biomédica. Lo anterior ayudaría a establecer sus alcances y limitaciones para integrarlos a la práctica docente en este campo, y, por otro lado, permitiría identificar oportunidades para mejorar su desempeño. En este artículo presentamos un procedimiento novedoso para evaluar la calidad de preguntas generadas a través de los modelos de lenguaje ChatGPT y Bard, que conformaron un instrumento de evaluación de 6 preguntas, y que fue aplicado a estudiantes de la carrera de Ingeniería Biomédica.

Modelos de Lenguaje

Los LLMs han permitido generar agentes que proporcionan un estilo de interacción conversacional, similar al de las personas [4][5] además de que no limitan su respuesta a dominios específicos de conocimiento.

Un modelo de lenguaje o LLM, tal como el de los agentes ChatGPT y Bard, resultan de pre-entrenar una red neuronal con una gran cantidad de datos públicos [6][7][8]. El modelo resultante es una función matemática que representa la distribución de probabilidad de secuencias de palabras. Entonces, dada una secuencia de palabras, los modelos de lenguaje de estos agentes, generan el siguiente token de palabra dependiendo de la probabilidad estimada para las palabras anteriores. Este mecanismo subyacente de los LLMs de predecir la siguiente palabra es uno de los factores importantes que explican por qué los agentes basados en LLMs pueden generar respuestas diferentes para la misma pregunta, que suenan plausibles, pero que, además, podrían contener errores fácticos [6]. Tales modelos son perfeccionados mediante un proceso conocido como aprendizaje por refuerzo a partir de la retroalimentación humana (RLHF), que consiste en entrenadores de IA humanos que brindan recompensas al modelo para ayudarlo a corregir sus errores [6][7][8].

Adopción de los Modelos de Lenguaje

El ChatGPT (del inglés, *"Generative Pre-trained Transformer"*), fue lanzado por el laboratorio de investigación de inteligencia artificial, OpenAI, en noviembre de 2022. Tal evento incrementó rápidamente el interés de la sociedad por la Inteligencia Artificial ya que el volumen de tráfico en el Internet por buscar la palabra "IA" y "ChatGPT" se triplicó, pasando de 7.9 millones de búsquedas en noviembre del 2022 a más de 30.4 millones a principios del 2023 [9]. La adopción del ChatGPT ha sido principalmente por la población adulta joven de 18 a 44 años [10], lo que incluye estudiantes. Por ejemplo, dentro de la enseñanza en el área del cuidado de la salud, se ha propuesto su uso para mejorar el aprendizaje personalizado como tutor mediante la resolución de dudas, y elaboración de resúmenes [3][11]. Recientemente, en marzo del 2023, Google lanzó el agente conversacional Bard, considerado el competidor directo de ChatGPT. Aun con el poco tiempo que tiene disponible para accederse por los usuarios, se puede identificar el interés por estudiar su utilidad en el campo de la salud [12]. Este interés radica principalmente, por sus similitudes en las tecnologías de IA que utilizan, pero son modelos que resultaron de un entrenamiento con diferentes fuentes de datos públicos, pero de diferentes tamaños. El modelo de ChatGPT fue entrenado en 2021 con una fuente de datos de texto

masiva, mayor que la usada para entrenar el modelo de Bard, mientras que Bard fue entrenado con datos recientes obtenidos exclusivamente de la Web y es capaz de buscar respuestas en Internet en tiempo real.

El interés en estos agentes conversacionales incluye consultarles sobre temas especializados del mundo académico para facilitar el proceso de enseñanza-aprendizaje [3][12][13], ya que son capaces de dialogar de forma escrita, dando respuestas que potencialmente son consistentes y contextualmente apropiadas a las preguntas del usuario. Sin embargo, es necesario tener en cuenta que estas herramientas no pueden interpretar ni comprender a profundidad el contenido de un diálogo, por lo que podrían generar información incorrecta [14]. Un problema reconocido de la tecnología LLM es su tendencia a producir información que no se basa en sus datos de entrenamiento, conocida como "alucinación" [15]. En consecuencia, la información que generan podría erróneamente percibirse como confiable por lectores no especializados, tal como estudiantes.

Trabajos previos

Para utilizar de forma confiable a los agentes conversacionales basados en LLMs, se han formulado guías para integrarlos a la educación superior, que sugieren verificar los hechos o conocimiento que proporcionan, utilizar tal conocimiento con un juicio crítico, verificando fuentes alternas de información [16]. Otros estudios, han evaluado su utilidad para asistir en actividades y áreas académicas específicas. En esta dirección, algunos estudios han evaluado el desempeño del ChatGPT para responder preguntas de exámenes profesionales de diversas especialidades, tal como el examen de la Barra de Abogados de EE. UU [17][18] y de Licencia Médica de EE. UU. (USMLE) [19][20] con una exactitud superior al 57.6 %. Por otro lado, los resultados en términos de la precisión de ChatGPT puede variar entre disciplinas o temas de una misma disciplina. Por ejemplo, mientras que el ChatGPT resultó eficaz para responder preguntas de conocimiento de primer y segundo orden sobre microbiología con una precisión del 80% [4], su rendimiento para responder preguntas clínicas en el campo de la enfermedad glomerular es deficiente (<=60 %)[21]. Lo anterior puede deberse a que los datos de entrenamiento para generar el modelo ChatGPT carecieron de datos ejemplos suficientes en esos temas [3].

Algunos estudios han investigado el uso de agentes conversacionales basados en LLM para generar instrumentos de evaluación, ver Tabla 1. Cheung utilizó ChatGPT Plus para generar 50 preguntas de opción múltiple sobre temas de posgrado de medicina interna y cirugía [22]. Su estudio evaluó la idoneidad, claridad, especificidad, relevancia y poder discriminativo de las preguntas, comparándolas con 50 preguntas generadas por expertos humanos. Los resultados mostraron que ChatGPT Plus generó preguntas con calidad comparable a las del experto humano [22]. Kumar usó ChatGPT para generar seis preguntas de opción múltiple sobre otosclerosis para evaluar diferentes niveles de aprendizaje con diferentes grados de dificultad, pero sin realizar una evaluación formal [23]. Elkins utilizó InstructGPT para generar 612 preguntas abiertas de nivel secundaria sobre temas de biología y aprendizaje automático, utilizando la taxonomía de *Bloom* para evaluar diferentes niveles de aprendizaje [24]. Once expertos evaluaron la calidad de las preguntas en términos de relevancia, gramática y utilidad, con resultados favorables [24]. Yuan utilizó GPT-3 para crear cinco preguntas abiertas para estudiantes de primaria y secundaria sobre diversos temas [25]. Su objetivo era evaluar un método para seleccionar preguntas de calidad entre un conjunto de preguntas generadas con IA. Se realizó evaluación automática contra preguntas de referencia y evaluación por 87 expertos demostrando la eficacia de su método [25]. Finalmente, Xiao utilizó ChatGPT y GPT-2 para realizar 20 preguntas de opción múltiple sobre inglés nivel secundaria [26]. Se realizó evaluación automática y por 373 estudiantes. Sus resultados indicaron que el material generado por IA puede superar en calidad al material escrito por humanos [26].

Este trabajo considera que agentes conversacionales, como ChatGPT y Bard, son una poderosa herramienta para simplificar y mejorar la eficiencia de tareas de los educadores, como las relacionadas con resumir y traducir información [13], organizar y generar material y corregir textos [27][3]. Sin embargo, no hay evidencia de su desempeño para generar exámenes de conocimiento en temas de Ingeniería Biomédica. Consecuentemente, es importante determinar cuáles son las limitaciones y posibles sesgos de utilizar agentes conversacionales basados en LLMs para apoyar tales tareas académicas. Sin embargo, las investigaciones son escasas y limitadas a dominios educativos más generales que requieren niveles de pensamiento básicos para su resolución, tal como usar los LLMs para generar exámenes sobre la comprensión de textos [28]. La adquisición de conocimiento puede evaluarse a diferentes niveles, desde recordar y comprender hasta evaluar y crear de acuerdo con la taxonomía de Bloom [29]. Aquí nos enfocamos en la aplicación de conocimiento que se encuentra en un nivel superior a los estudios previos. Evaluar la aplicación de conocimiento de los estudiantes permite evaluar su capacidad no solo de comprender información sino de usar el conocimiento aprendido en situaciones nuevas [29]. Este nivel de aprendizaje es fundamental para garantizar que los estudiantes podrán utilizar su conocimiento para la solución de problemas reales. Lo anterior, nos motivó a realizar esta investigación con el siguiente propósito.

Objetivo

El objetivo de esta investigación fue analizar la calidad de ChatGPT y Bard para crear instrumentos de evaluación en temas de la Ingeniería Biomédica. Debido a las limitaciones de la tecnología LLM, en este trabajo se utiliza un enfoque crítico para evaluar y comparar la calidad en términos de validez, relevancia, claridad, dificultad y capacidad de discriminación de las preguntas generadas por los agentes conversacionales para la evaluación de aplicación de conocimiento sobre medición de señales bioeléctricas en comparación con las de un experto humano.

MATERIALES Y MÉTODOS

Desarrollo del Instrumento de Evaluación

Un experto humano (agente E) y dos agentes conversacionales (agentes B y C) generaron seis potenciales preguntas cada quién para integrar un instrumento de evaluación para estudiantes de octavo semestre de ingeniería biomédica al final de un curso sobre mediciones biomédicas impartido por el agente E. Previo a la generación de todas las preguntas, el agente E definió que el instrumento de evaluación estuviera integrado por preguntas de forma equitativa tanto de agentes como de temas a un nivel de aplicación de conocimiento dentro de la taxonomía de Bloom [29]. Los temas evaluados fueron electrocardiografía, electromiografía, electroencefalografía, electro-oculografía, electrorretinografía y magneto cardiografía.

El agente E utilizó Bard como agente B y ChatGPT-3 como agente C debido a que Bard y ChatGPT-3 son los dos agentes más utilizados en mayo de 2023 con acceso gratuito, facilitando su uso para esta investigación y la generalización de nuestros resultados. Bard y ChatGPT-3 son agentes ampliamente investigados, con buena reputación, y capacidades comparables debido a la similitud de sus tecnologías.

"Create an exam for undergraduate students of biomedical engineering on the applying levels of thinking using the Bloom's taxonomy. Include one questions for the Each of the following topics Electrocardiography, electromyography, electroencephalography, electrooculography, electroretinography, and magnetocardiography. Each question must have 3 answer options with only one correct answer. Highlight the right answer for each question."

FIGURA 1. Ejemplo de un mensaje de ChatGPT-3 para generar preguntas a nivel de aplicación sobre la interfaz electrodo-electrolito en señales bioeléctricas.

TABLA 1. Trabajos previos sobre LLMs para generación de instrumentos de evaluación de conocimiento.

Dominio / Artículo	Cheung 2023[22]	Kumar 2023[23]	Elkins 2023[24]	Yuan 2023[25]	Xiao 2023[26]
Modelo de Grandes Lenguajes	ChatGPT plus	ChatGPT	InstructGPT	GPT-3	ChatGPT y GPT-2
Tipo de preguntas	Opción múltiple	Opción múltiple	Abiertas	Abiertas	Opción múltiple
Número de preguntas	50 generadas por IA y 50 por dos expertos humanos	6 preguntas generadas por IA	612 preguntas generadas por IA	5 preguntas	20 preguntas
Nivel y disciplina	Posgrado en Medicina	Medicina	secundaria	Nivel primaria y secundaria	Inglés nivel secundaria
Temas específicos	Medicina interna y cirugía	Otoesclerosis	Machine Learning y Biología	Temás diversos y de literatura infantil	Inglés
Nivel de aprendizaje evaluado	No especificado	Memoria, comprensión, afectivo y diferente nivel de dificultad.	Recordar, comprender, aplicar, analizar, evaluar y crear según taxonomía de Bloom en niveles principiantes, intermedios y avanzados	Comprensión de lectura	Comprensión de lectura
Evaluación	Comparación IA contra experto humano por cinco expertos mediante un formulario estandarizado sobre los siguientes dominios: Idoneidad, claridad y especificidad, relevancia, poder discriminativo, idoneidad para posgrado	Ninguna	Evaluación por 11 expertos sobre relevancia, gramática, responsabilidad, adherencia y utilidad	Evaluación automática contra referencias en términos de similitud y equivalencia semántica. Evaluación por 87 expertos de corrección gramatical, ofensividad, claridad, relevancia, importancia, especificidad, y capacidad de respuesta	Evaluación automática de diversidad y legibilidad. Evaluación por 373 estudiantes de legibilidad, corrección, coherencia, compromiso, calidad general, calidad relativa, probabilidad de generación por IA, coherencia temática, idoneidad, coincidencia de contenido, utilidad, idoneidad, similitud,
Resultados	La IA puede generar preguntas de posgrado en medicina con calidad comparable a las de un experto. La relevancia de las preguntas de la IA fue menor (p = 0.04).	La IA puede generar preguntas de opción múltiple con diferentes niveles de aprendizaje, y niveles de dificultad.	Las preguntas generadas con IA son de alta calidad (>67 %, k Cohen >0.53) en términos de relevancia, gramática, responsabilidad, adherencia, y suficientemente útiles.	Se demostró la efectividad del método para seleccionar preguntas de alta calidad entre las generadas por la IA.	El material generado por IA puede superar la calidad del material escrito por humanos.

Como se mencionó ambos fueron entrenados con un conjunto de datos público masivo, ambos tienen capacidad para escribir contenido y responder preguntas. Estos dos agentes fueron elegidos debido a su representatividad en la generación de lenguaje natural con el objetivo de simplificar el proceso de comparación y hacerlo manejable para realizar un análisis profundo y detallado tomando en consideración limitaciones de tiempo y presupuesto. En la Figura 1 se muestra un ejemplo de solicitud para que ChatGPT-3 genere sus preguntas a nivel de aplicación de conocimiento sobre los temas seleccionados. Se realizó una solicitud similar para Bard.

El agente E analizó la calidad de cada pregunta para seleccionar aquellas que pudieran integrar el instrumento de evaluación de final respectando el principio de equidad y en orden aleatorio.

Aplicación y Evaluación del Instrumento

Posteriormente un grupo de expertos evaluaron atributos de calidad de las seis preguntas respecto a claridad, validez, nivel de pensamiento de acuerdo con la Taxonomía de Bloom, dificultad, validez y relevancia. Este grupo de expertos estaba compuesto por el agente E y dos expertos adicionales con al menos cinco años de experiencia docente en los temas del curso. Para realizar esta evaluación utilizaron un formulario estandarizado en Word. El formulario contenía una explicación de apoyo para cada atributo a evaluar. La claridad, relevancia y validez de las preguntas se evaluaron mediante una variable dicotómica (sí/no). Además, clasificaron cada pregunta en un nivel de pensamiento (aplicación, análisis o evaluación) y especificaron un grado de dificultad (es decir, bajo, medio o alto). La claridad, validez, nivel de pensamiento, dificultad, validez y relevancia se evaluaron midiendo la consistencia en las respuestas de los expertos humanos a estos aspectos.

El instrumento de evaluación se tradujo y aplicó en español. Las preguntas fueron registradas en la plataforma de aprendizaje Moodle para crear el banco de preguntas del instrumento. Se reclutó a todos los estudiantes (es decir, 39) del último año de Ingeniería Biomédica de la Universidad, hispanohablantes que tomaron el curso sobre Mediciones Biomédicas. Sólo un estudiante no aceptó participar, quedando treinta y ocho participantes que dieron su consentimiento para utilizar sus resultados. Estos participantes accedieron a la plataforma para contestar el examen en línea al mismo tiempo y desde las instalaciones de la Universidad. El instrumento fue configurado para que cada pregunta se presentara de forma individual y secuencial acompañada de una pregunta dicotómica para que los estudiantes calificaran la claridad de las preguntas. El tiempo máximo para la resolución del instrumento de evaluación fue de 2 minutos por pregunta.

Análisis de Estadístico

Se realizó estadística descriptiva general de los resultados obtenidos de los participantes. Se estimó la consistencia entre los evaluadores utilizando el índice kappa de Fleiss (k), ver ecuación (1).

$$k = \frac{P_o - P_e}{1 - P_e} \tag{1}$$

Donde P_o es la proporción observada de acuerdo entre los evaluadores y P_e es la proporción esperada de acuerdo debido al azar. Para cada pregunta se calcularon índices de dificultad y discriminación. La claridad evaluada por los estudiantes (c) se calculó como la proporción de estudiantes (x) que consideraron clara la pregunta, ver ecuación (2).

$$c = \frac{n}{N} \tag{2}$$

Donde N es el número total de respuestas.

Para evaluar el índice de discriminación, la muestra de estudiantes se dividió en tres grupos equitativos (33 %)

en función de su desempeño para comparar los resultados entre el grupo con mayor y menor desempeño. Se realizó un análisis Rasch con el software gratuito Jamovi [30], ver ecuación (3).

$$P(X_{ni}=1)=\frac{e^{(d_i-h_n)}}{1+e^{(d_i-h_n)}} \quad (3)$$

Donde P es la probabilidad de que un individuo n responda correctamente la pregunta i, d_i es la dificultad de la pregunta i y h_n es la habilidad del individuo n.

Esto se utilizó para calcular el valor p para el ajuste del modelo, la matriz de correlación de todas las preguntas del examen y el mapa de habilidades de *Wright*, ver ecuación (4).

$$d_i-h_n=ln\left(\frac{P(X_{ni}=1)}{1-P(X_{ni}=1)}\right) \quad (4)$$

Las asociaciones error-claridad y agente generador con dificultad y claridad se evaluaron mediante Chi-cuadrado. Las diferencias entre agentes en cuanto a claridad, dificultad y discriminación se analizaron mediante pruebas de Kruskal Wallis (H), ver ecuación (5).

$$H=\frac{12}{N(N+1)}\sum\frac{R_j^2}{n_j}-3(N+1) \quad (5)$$

Donde N es el número total de observaciones, R_j es la suma de rangos del grupo j y n_j es el tamaño del grupo.

RESULTADOS Y DISCUSIÓN

El agente E eliminó cuatro preguntas generadas por el agente C por problemas de validez y tres preguntas del agente B por problemas de repetición. De esta forma, se seleccionaron dos preguntas por agente, utilizando las preguntas del agente E necesarias para completar la evaluación de los seis temas propuestos respetando el principio de equidad. El diseño del instrumento de evaluación final según los agentes quedó BECBCE. Los expertos mostraron un acuerdo perfecto calificando las seis preguntas del instrumento de evaluación como claras, así como válidas y relevantes. Los expertos mostraron un acuerdo razonable (kappa=0.22) y casi perfecto (kappa=0.83) en cuanto a su asignación de nivel de pensamiento y dificultad de las preguntas del instrumento de evaluación, ver Tabla 2. De forma general el instrumento integró dos preguntas de cada uno de los tres niveles de dificultad. La mayoría de las preguntas fueron interpretadas como preguntas de análisis de acuerdo con los expertos a excepción de la pregunta dos que fue considerada como de aplicación por dos expertos. Así, el agente E fue el único en generar una pregunta de acuerdo con el nivel de pensamiento deseado. La calificación promedio obtenida fue de 7.24, con una desviación estándar de 2.40, en un rango de puntuación total de 1.67 a 10 puntos.

TABLA 2. Análisis de expertos de las preguntas del instrumento de evaluación.

Pregunta	Agente	Nivel de pensamiento			Dificultad		
		E	E1	E2	E	E1	E2
Q1	B	2	2	2	m	m	m
Q2	E	1	3	1	m	a	m
Q3	C	2	2	2	a	a	a
Q4	B	2	2	2	b	b	b
Q5	C	1	2	2	b	b	b
Q6	E	1	2	2	a	a	a

E: agente E, E1: evaluador 1, E2: evaluador 2. Para nivel de pensamiento: 1 = aplicación, 2 = análisis, 3 = evaluación. Para dificultad: b = baja, m = media, a = alta.

El análisis de Rasch reveló un valor p para el ajuste del modelo de 0.272. La matriz de correlación no mostró correlaciones fuertes entre preguntas, ver Tabla 3. La correlación más alta (r=-0.37) se encontró entre las preguntas Q1, Q5 y Q6 que corresponden a preguntas de electrocardiografía, electroretinografía y magnetocardiografía de tres agentes diferentes.

TABLA 3. Matriz de correlación del análisis Rasch de preguntas del instrumento de evaluacióninstrumento de evaluación.

	Q1	Q2	Q3	Q4	Q5
Q2	-0.26	-			
Q3	-0.15	-0.27	-		
Q4	0.16	-0.22	0.08	-	
Q5	-0.37	0.19	-0.14	-0.27	-
Q6	-0.37	0.28	0.06	-0.13	-0.07

Las preguntas del instrumento presentaron un índice de dificultad promedio de 0.72, con una desviación estándar de 0.09, ver Tabla 4. De acuerdo con este índice, la mayoría de las preguntas (N=4) pueden ser consideradas fáciles con un índice de dificultad >0.7. Sin embargo, no siempre el índice de dificultad calculado en base a la respuesta de los alumnos correspondió con la calificación de dificultad asignada por los expertos. Por ejemplo, mientras que las pregunta Q6 calificada como de alta dificultad por los expertos obtuvo uno de los menores índices de dificultad (0.74).

Los estudiantes asignaron una puntuación de claridad a las preguntas en promedio de 97 % con una desviación estándar de 3 %, ver Tabla 4. Dos preguntas obtuvieron una calificación perfecta de claridad correspondientes a las preguntas sobre temas de electrocardiografía y electro oculografía generadas por el agente B.

En la evaluación del índice de discriminación, el grupo de estudiantes con el peor desempeño estuvo conformado por 12 estudiantes que obtuvieron un puntaje menor o igual a 5.00 (32 %), mientras que el grupo con el mejor desempeño por 11 estudiantes (29 %) que obtuvieron una calificación de 10. La pregunta con el índice de dificultad mínimo (0.87) fue generada por el agente C, mientras que la de mayor índice de dificultad fue generada por el agente E. El índice de discriminación promedio fue de 0.31 con una desviación estándar de 0.13. Cinco preguntas (83 %) presentaron un índice de discriminación adecuado (> 0.2). La pregunta con menor índice de dificultad generada por el agente C fue la que también presentó el menor índice de discriminación (0.37).

TABLA 4. Índice de dificultad y discriminación de las preguntas del instrumento de evaluación.

Pregunta	Claridad	Índice de dificultad	Índice de discriminación
Q1	1.00	0.74	0.21
Q2	0.97	0.61	0.47
Q3	0.95	0.66	0.37
Q4	1.00	0.74	0.32
Q5	0.92	0.87	0.11
Q6	0.97	0.74	0.37

El mapa de competencias de Wright del análisis Rasch mostró una distribución de competencias de los estudiantes con mayor agrupación por debajo de la media y dos estudiantes sobresalientes, ver Figura 2. El mapa de Wright también mostro una distribución de dificultad de las preguntas con mayor agrupación por encima de la media, con un efecto de techo debido a los once estudiantes que obtuvieron un puntaje perfecto. También se observó redundancia en las preguntas Q1, Q4 y Q6 que corresponden a los temas de electrocardiografía, electro-oculografía y magneto cardiografía. Dos de estas preguntas fueron generadas por el agente B y una por el agente E.

No se encontraron asociaciones significativas (p>0.06) error-claridad, ni agente generador con la dificultad y claridad. No se encontraron diferencias estadísticamente significativas (p>0.08) por agentes en claridad, dificultad y discriminación mediante pruebas de Kruskal Wallis. No se utilizaron pruebas paramétricas debido al número reducido de preguntas que fueron incluidas en el instrumento de evaluación final. La versión final del instrumento de evaluación se muestra en la Tabla 5. Este es el primer trabajo que evalúa mediante un análisis crítico la calidad de los agentes conversacionales en la evaluación de temas relacionados con mediciones biomédicas.

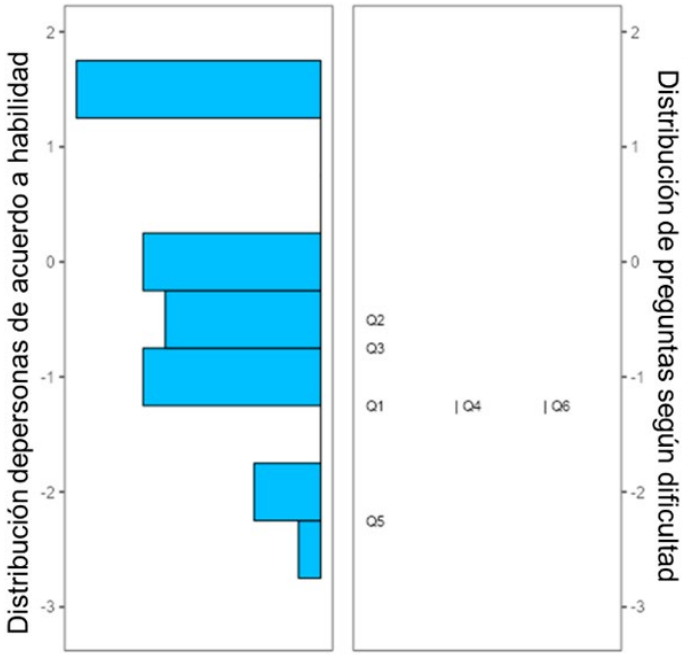

FIGURA 2. **Mapa de Wrigth de análisis Rasch del instrumento de evaluación**

Esto es particularmente destacable para el caso del agente Bard del que no se cuenta con experimentos similares documentados tal vez debido a su más reciente lanzamiento. De forma similar es el primer experimento en realizar una evaluación a un nivel de pensamiento superior al de la simple comprensión de información. La falta de consenso en cuanto a la evaluación de nivel de pensamiento de las preguntas por parte de los expertos puede demostrar la necesidad de revisar el establecimiento de un marco de referencia común entre ellos. Sin embargo, puede identificarse un consenso en la existencia de una diferencia entre el nivel de pensamiento deseado y el generado por los agentes virtuales, tal vez debido a sus limitaciones en interpretación y comprensión [14]. Los problemas de relevancia y originalidad detectados en las preguntas generadas por los agentes virtuales sugieren una limitación para poder utilizar estas herramientas sin supervisión experta.

El análisis de Rasch mostró elevada competencia de los evaluados y baja complejidad de las preguntas. Se puede afirmar que las preguntas del agente E presentaron mayor índice de dificultad (67 %) y discriminación (42 %). Esto pudiera deberse a que este agente fue el que impartió el curso y conocía el contexto de fortalezas y debilidades del grupo. Mientras tanto el agente B generó preguntas redundantes con los menores índices de dificultad (0.74). Por otro lado, el agente C generó preguntas con excelente claridad. De forma interesante, el índice de dificultad del análisis Rasch no siempre guardó relación con el nivel de dificultad asignado por los evaluadores.

Esto revela la importancia del análisis de la dificultad de las preguntas no solo en base al juicio de los expertos sino en el desempeño de los estudiantes. Por otro lado, la agrupación de estudiantes en el extremo superior de desempeño habla de un efecto de techo del instrumento de evaluación.

La ingeniería biomédica es un campo interdisciplinario donde la educación es un componente crucial para la preparación de los futuros ingenieros. De esta manera, los avances en metodologías de aprendizaje y estrategias de evaluación son temas relevantes. Artículos como el presente pueden contribuir a mejorar la calidad de la formación en el área. Por otro lado, la evolución de las formas de aprendizaje ya está impactando la educación en ingeniería biomédica, por lo que los artículos sobre estos temas pueden ayudar a los profesionales a mantenerse actualizados y enfrentar nuevos desafíos. El objetivo principal de nuestra investigación fue evaluar la calidad de agentes conversacionales en el contexto de la medición de señales bioeléctricas, lo cual es fundamental en ingeniería biomédica. Así nuestro artículo es una contribución original a la evaluación de la calidad de los agentes conversacionales a un nivel de pensamiento más significativo que la simple comprensión de conceptos en el área de la medición de señales biomédicas. Creemos importante discutir la necesidad de desarrollar habilidades en estudiantes y profesionales sobre el manejo de agentes conversacionales.

TABLA 5. Instrumento de evaluación final.

Q1: ¿Cuál de las siguientes es la derivación más utilizada en electrocardiografía? •Derivación I •Derivación II •Derivación III
Q2: Está realizando el análisis de un estudio de electromiografía de superficie de un paciente ¿Cuál es el orden correcto para procesar la señal de electromiografía de superficie? •Procesamiento de línea base, rectificación y suavizado •Rectificación, suavizado y procesamiento de línea base •Suavizado, procesamiento de línea base y rectificación
Q3: El registro de EEG de un paciente muestra un ritmo alfa prominente en la región occipital. ¿Qué estado del cerebro es más probable que se asocie con este hallazgo? •Despierto y relajado •Sueño profundo •Alta actividad cognitiva
Q4: ¿Cuál de los siguientes es el electrodo más común utilizado en electrooculografía? •Electrodo de superficie •Electrodo de aguja •Electrodo de alambre
Q5: Está realizando una electrorretinografía en un paciente. ¿Qué configuración de colocación de electrodos se usa comúnmente para registrar las respuestas eléctricas de la retina? •Fpz - Oz •Cz - Pz •Electrodo activo en la córnea y electrodo de referencia en la piel
Q6: Está realizando un estudio de magneto cardiografía a un paciente. ¿Cuáles serían la amplitud y frecuencia máxima de la señal esperados? •5 pT y 100 Hz •5 nT y 1000 Hz •5 fT y 10 Hz

Por otro lado, aunque el número de preguntas incluido en nuestro instrumento de evaluación puede ser una limitante, consideramos que el instrumento actual permitió cumplir con el objetivo de evaluar la calidad del instrumento de evaluación creado con la ayuda de agentes conversacionales en todos los términos propuestos: nivel de pensamiento, validez, relevancia, claridad, dificultad y capacidad de discriminación. Además, los métodos que utilizamos para evaluar la calidad del cuestionario, como la índice kappa (k) para evaluar la concordancia entre los expertos y el análisis de Rasch para determinar la consistencia de los resultados y de los estudiantes, son robustos ya que se centran en analizar la calidad del instrumento en lugar del mero número de pruebas. Sin duda, aumentar el número de preguntas contribuiría a la fiabilidad matemática del estudio. Pero el incremento en el número de preguntas podría afectar las respuestas de los expertos y estudiantes al afectar su calidad y tasa de participación. Por lo tanto, consideramos que, en su versión actual, el instrumento permitió cumplir con los objetivos del estudio y realizar una evaluación eficiente sin comprometer la calidad de los datos.

CONCLUSIONES

Los agentes conversacionales basados en LLMSs tienen capacidad para evaluar la aplicación de conocimiento en medición de señales bioeléctricas. Aunque no se encontraron diferencias de calidad en términos de claridad, dificultad y discriminación en comparación con un experto, se encontraron problemas de validez y originalidad que requieren supervisión de un experto.

Por lo tanto, es necesario realizar más pruebas y estudios para determinar la calidad de la información. Si bien, estas herramientas tienen potencial, la evidencia sustancial que respalda su uso en entornos académicos es fundamental. La alfabetización en estas herramientas en los programas de formación educativa es fundamental. Los docentes deben aprender a usar estas herramientas y comprender claramente sus limitaciones y los objetivos que quieren lograr al usarlas [3]. También es esencial dar a los académicos la autoridad y la independencia para utilizarlos [11]. Sin embargo, es importante que los expertos en el campo correspondiente validen el contenido de las preguntas generadas por estas herramientas para asegurar su calidad y

garantizar que cumpla con los estándares académicos requeridos.

DECLARACIÓN ÉTICA

A todos los participantes en este estudio se les solicitó su consentimiento para participar en esta investigación de manera voluntaria previo a la recopilación y análisis de cualquier información. Los participantes que no aceptaron participar fueron excluidos de este estudio.

AGRADECIMIENTOS

Agradecimiento a todos los participantes que ayudaron a esta investigación.

CONTRIBUCIÓN DE AUTORES

A.O.R.M. participó en la curación de datos, realizó investigación y validó resultados, asimismo obtuvo recursos para el desarrollo del proyecto, participó en las diferentes etapas del desarrollo del manuscrito. A.M.P. conceptualizó el proyecto, realizó análisis formales, desarrolló la metodología y validó resultados, participó en la supervisión general del proyecto y participó en las diferentes etapas de la escritura del manuscrito. A.I.P.S. conceptualizó el proyecto, participó en la curación de datos e implementó software especializado, desarrolló análisis formales, validó y visualizó los resultados, realizó investigación y desarrolló la metodología para el proyecto, obtuvo recursos para el desarrollo del proyecto, y se encargó de la administración y supervisión general, participó en las diferentes etapas de la escritura del manuscrito. M.D.R.U. participó en la obtención de fondos y recursos para el desarrollo del proyecto, desarrolló la metodología, la visualización y validación de los resultados, y participó en las diferentes etapas del desarrollo del manuscrito. M.C.A.R. participó en la obtención de recursos para el desarrollo del proyecto, realizó investigación y validó resultados, y participó en las diferentes etapas del desarrollo del manuscrito. Todos los autores aprobaron la versión final del manuscrito.

REFERENCES

[1] OpenAI, "ChatGPT" [Large language model]. OpenAI. https://chat.openai.com/chat (consultado en 2023).

[2] Google, "Bard" [Large Language Model]. 2023. https://bard.google.com/chat (consultado en 2023).

[3] M. Sallam, "ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns," Healthcare, vol. 11, no. 6, art. no. 887, mar. 2023, doi: https://doi.org/10.3390/healthcare11060887

[4] M. McTear, Z. Callejas, D. Griol, "The conversational interface: Talking to smart devices," in The Conversational Interface, Switzerland: Springer International Publishing, 2016, doi: https://doi.org/10.1007/978-3-319-32967-3

[5] M. M. E. Van Pinxteren, M. Pluymaekers, J. G. A. M. Lemmink, "Human-like communication in conversational agents: a literature review and research agenda," J. Serv. Manag., vol. 31, no. 2, pp. 203-225, 2020, doi: https://doi.org/10.1108/JOSM-06-2019-0175

[6] J. Manyika, "An overview of Bard: an early experiment with generative AI," Google. 2023. [En línea]. Disponible en: https://ai.google/static/documents/google-about-bard.pdf

[7] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, et al., "Language models are few-shot learners," 2020, arXiv:2005.14165, doi: https://doi.org/10.48550/arXiv.2005.14165

[8] O. Vinyals, Q. Le, "A Neural Conversational Model," 2015, arXiv: 1506.05869, doi : https://doi.org/10.48550/arXiv.1506.05869

[9] T. Bianchi, "Global search volume for 'ChatGPT API', 'AI API' keywords 2022-2023," Statista. Disponible en: https://www.statista.com/statistics/1398265/chatgpt-ai-api-keywords-search-volume/ (consultado el 26 de octubre de 2023).

[10] A. Petrosyan, "ChatGPT and cyber crime - Statistics & Facts," Statista. Disponible en: https://www.statista.com/topics/10818/chatgpt-and-cyber-crime/#topicOverview (consultado el 26 de octubre de 2023).

[11] R. S. D'Amico, T. G. White, H. A. Shah, D. J. Langer, "I Asked a ChatGPT to Write an Editorial About How We Can Incorporate Chatbots Into Neurosurgical Research and Patient Care...," Neurosurgery, vol. 92, no. 4, pp. 663-664, abr. 2023, doi: https://doi.org/10.1227/neu.0000000000002414

[12] F. Y. Al-Ashwal, M. Zawiah, L. Gharaibeh, R. Abu-Farha, A. N. Bitar, "Evaluating the Sensitivity, Specificity, and Accuracy of ChatGPT-3.5, ChatGPT-4, Bing AI, and Bard Against Conventional Drug-Drug Interactions Clinical Tools," Drug Healthc. Patient. Saf., vol. 15, pp. 137-147, sep. 2023, doi: https://doi.org/10.2147/dhps.s425858

[13] H. Yang, "How I use ChatGPT responsibly in my teaching," Nature, apr. 2023, doi: https://doi.org/10.1038/d41586-023-01026-9

[14] S. Ariyaratne, K. P. Iyengar, N. Nischal, N. Chitti Babu, R. Botchu, "A comparison of ChatGPT-generated articles with human-written articles," Skeletal Radiol., vol. 52, no. 9, pp. 1755-1758, sep. 2023, doi: https://doi.org/10.1007/s00256-023-04340-5

[15] G. Eysenbach, "The Role of ChatGPT, Generative Language Models, and Artificial Intelligence in Medical Education: A Conversation With ChatGPT and a Call for Papers," JMIR Med. Educ., vol. 9, no. 4, art. no. e46885, mar. 2023, doi: https://doi.org/10.2196/46885

[16] D. De Silva, N. Mills, M. El-Ayoubi, M. Manic, D. Alahakoon, "ChatGPT and Generative AI Guidelines for Addressing Academic Integrity and Augmenting Pre-Existing Chatbots," in 2023 IEEE International Conference on Industrial Technology (ICIT), Orlando, FL, USA, 2023, pp. 1-6. doi: https://doi.org/10.1109/ICIT58465.2023.10143123

[17] M. J. Bommarito, D. M. Katz, "GPT Takes the Bar Exam," SSRN Electron. J., pp. 1-13, 2023, doi: https://dx.doi.org/10.2139/ssrn.4314839

[18] J. Bommarito, M. J. Bommarito, J. Katz, D. M. Katz, "Gpt as Knowledge Worker: A Zero-Shot Evaluation of (AI)CPA Capabilities," SSRN Electron. J., 2023, doi: https://dx.doi.org/10.2139/ssrn.4322372

[19] A. Gilson, C. W. Safranek, T. Huang, V. Socrates, L. Chi, R. A. Taylor, D. Chartash, "How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment," JMIR Med. Educ., vol. 9, art. no. e45312, 2023, doi: https://doi.org/10.2196/45312

[20] T. H. Kung, M. Cheathem, A. Medenilla, C. Sillos, et al., "Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models," PLOS Digit. Health, vol. 2, no. 2, art. no. e0000198, 2023, doi: https://doi.org/10.1371/journal.pdig.0000198

[21] J. Miao, C. Thongprayoon, W. Cheungpasitporn, "Assessing the Accuracy of ChatGPT on Core Questions in Glomerular Disease," Kidney Int. Rep., vol. 8, no. 8, pp. 1657-1659, 2023, doi: https://doi.org/10.1016/j.ekir.2023.05.014

[22] B. H. H. Cheung, G. K. K. Lau, G. T. C. Wong, E. Y. P. Lee, et al., "ChatGPT versus human in generating medical graduate exam multiple choice questions—A multinational prospective study (Hong Kong S.A.R., Singapore, Ireland, and the United Kingdom)," PLoS One, vol. 18, no. 8, art. no. e0290691, ago. 2023, doi: https://doi.org/10.1371/journal.pone.0290691

[23] A. K. Khilnani, "Potential of Large Language Model (ChatGPT) in Constructing Multiple Choice Questions," GAIMS J. Med. Sci., vol. 3, no. 2, pp. 1-3, 2023. [En línea]. Disponible en: https://gjms.gaims.ac.in/index.php/gjms/article/view/71

[24] S. Elkins, E. Kochmar, I. Serban, J. C. K. Cheung, "How Useful Are Educational Questions Generated by Large Language Models?," in Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky. AIED 2023. Communications in Computer and Information Science, vol 1831, Tokyo, Japón, 2023, pp. 536-542, doi: https://doi.org/10.1007/978-3-031-36336-8_83

[25] X. Yuan, T. Wang, Y.-H. Wang, E. Fine, R. Abdelghani, H. Sauzéon, P.-Y. Oudeyer, "Selecting Better Samples from Pre-trained LLMs: A Case Study on Question Generation," in Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canadá, 2023, pp. 12952-12965, doi: https://doi.org/10.18653/v1/2023.findings-acl.820

[26] C. Xiao, S. X. Xu, K. Zhang, Y. Wang, L. Xia, "Evaluating Reading Comprehension Exercises Generated by LLMs: A Showcase of ChatGPT in Education Applications," in Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023), Toronto, Canadá, 2023, pp. 610-625, doi: https://doi.org/10.18653/v1/2023.bea-1.52

[27] S. Sedaghat, "Early applications of ChatGPT in medical practice, education and research," Clin. Med., vol. 23, no. 3, pp. 278-279, may. 2023, doi: https://doi.org/10.7861/clinmed.2023-0078

[28] R. Dijkstra, Z. Genç, S. Kayal, and J. Kamps, "Reading Comprehension Quiz Generation using Generative Pre-trained Transformers," in 4th International Workshop on Intelligent Textbooks, iTextbooks 2022, Durham, Reino Unido, 2022. [En línea]. Disponible en: https://hdl.handle.net/11245.1/a1109043-92d4-4c63-be33-6e238780d3b7

[29] L. W. Anderson, D. R. Krathwohl, P. W. Airasian, K. A. Cruikshank, et al., A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives. Nueva York: Pearson Education, 2001.

[30] Jamovi. (2023). [En línea]. Disponible en: https://www.jamovi.org