ORIGINAL

Comparación de las respuestas a preguntas sobre intoxicaciones generadas por sistemas de inteligencia artificial y las creadas por toxicólogos clínicos

Santiago Nogué-Xarau[1,*], José Ríos-Guillermo[2,*], Montserrat Amigó-Tadín[3] y Grupo de Trabajo de Toxicología de la Societat Catalana de Medicina d'Urgències i Emergències (SoCMUETox)

Objetivos. Formular preguntas sobre intoxicaciones a cuatro sistemas de inteligencia artificial (IA) y a cuatro toxicólogos clínicos (TC) y constatar si un grupo de observadores es capaz de identificar el origen de las respuestas. Valorar la calidad del texto y el nivel de conocimientos ofrecidos por estas IA y compararlos con el de los TC.

Método. Se prepararon 10 preguntas de toxicología y se introdujeron en cuatro sistemas de IA (Copilot, Bard, LuzIA y ChatGPT). Se solicitó a cuatro TC que respondiesen a las mismas preguntas. Se consiguieron 24 observadores expertos en toxicología y se les remitió un cuestionario con 10 preguntas y cada una de ellas con una respuesta procedente de una IA y otra de un TC. Cada observador tenía que decidir la procedencia de las respuestas, valorar la calidad del texto y cuantificar el nivel de conocimientos sobre el tema.

Resultados. De las 240 respuestas que analizaron los observadores y que procedían de alguna IA, en 21 ocasiones (8,8%) opinaron que con certeza provenían de un TC, en 38 (15,8%) que procedían probablemente de un TC y en 13 (5,4%) reconocían que no podían establecer el origen de la respuesta. LuzIA y ChatGPT mostraron una mayor capacidad de engaño a los observadores, con diferencias significativas respecto a Bard (p = 0,036 y p = 0,041, respectivamente). Con relación a la calidad de los textos de las respuestas ofrecidas por las IA, la valoración de los observadores fue de excelente en el 38,8% de las ocasiones, con una diferencia significativa en favor de ChatGPT (61,3% de respuestas excelentes) respecto a Bard (34,4%, p < 0,001), LuzIA (31,7%, p < 0,001) y Copilot (26,3%, p < 0,001). Respecto a la percepción de conocimientos sobre el tema por parte de las IA, la puntuación media de fue de 7,23 (DE 1,57) sobre 10, obteniendo ChatGPT una puntuación de 8,03 (DE 1,26) que fue mayor a la obtenida por Luzia [7,02 (DE 1,63), p < 0,001], Bard [6,91 (1,64), p < 0,001] y Copilot [6,91 (1,46), p < 0,001].

Conclusiones. LuzIA y ChatGPT son sistemas de IA capaces de generar respuestas a preguntas de toxicología que, con frecuencia, parecen haber sido respondidas por un TC. La calidad de los textos generados y la percepción de conocimientos que ofrece ChatGPT es muy elevada.

Palabras clave: Inteligencia artificial. Toxicología clínica. Intoxicaciones. Test de Turing. Conocimiento.

Comparing answers of artificial intelligence systems and clinical toxicologists to questions about poisoning: Can their answers be distinguished?

Objectives. To present questions about poisoning to 4 artificial intelligence (AI) systems and 4 clinical toxicologists and determine whether readers can identify the source of the answers. To evaluate and compare text quality and level of knowledge found in the AI and toxicologists' responses.

Methods. Ten questions about toxicology were presented to the following AI systems: Copilot, Bard, Luzia, and ChatGPT. Four clinical toxicologists were asked to answer the same questions. Twenty-four recruited experts in toxicology were sent a pair of answers (1 from an AI system and one from a toxicologist) for each of the 10 questions. For each answer, the experts had to identify the source, evaluate text quality, and assess level of knowledge reflected. Quantitative variables were described as mean (SD) and qualitative ones as absolute frequency and proportion. A value of $P < .05$ was considered significant in all comparisons.

Results. Of the 240 evaluated AI answers, the expert evaluators thought that 21 (8.8%) and 38 (15.8%), respectively, were certainly or probably written by a toxicologist. The experts were unable to guess the source of 13 (5.4%) AI answers. Luzia and ChatGPT were better able to mislead the experts than Bard ($P = .036$ and $P = .041$, respectively). Text quality was judged excellent in 38.8% of the AI answers. ChatGPT text quality was rated highest (61.3% excellent) *vs* Bard (34.4%), Luzia (31.7%), and Copilot (26.3%) ($P < .001$, all comparisons). The average score for the level of knowledge perceived in the AI answers was 7.23 (1.57) out of 10. The highest average score was achieved by ChatGPT at 8.03 (1.26) *vs* Luzia (7.02 [1,63]), Bard (6.91 [1.64]), and Copilot (6.91 [1.46]) ($P < .001$, all comparisons).

Conclusions. Luzia and ChatGPT answers to the toxicology questions were often thought to resemble those of clinical toxicologists. ChatGPT answers were judged to be very well-written and reflect a very high level of knowledge.

Keywords: Artificial intelligence. Clinical toxicology. Poisoning. Turing test. Knowledge.

DOI: 10.55633/s3me/082.2024

*Estos autores han participado en igual medida en el manuscrito y deben ser considerados como primer autor.

Filiación de los autores:
[1]Fundación Española de Toxicología Clínica, Barcelona, España.
[2]Servicio de Farmacología Clínica. Hospital Clínic de Barcelona, España.
[3]Servicio de Urgencias, Hospital Clínic de Barcelona, España.

Contribución de los autores:
Todos los autores han confirmado su autoría en el documento de responsabilidades del autor, acuerdo de publicación y cesión de derechos a EMERGENCIAS.

Autor para correspondencia:
Santiago Nogué-Xarau
Fundación Española de Toxicología Clínica

Correo electrónico:
snoguex@gmail.com

Información del artículo:
Recibido: 23-6-2024
Aceptado: 24-7-2024
Online: 10-9-2024

Editor responsable:
Agustín Julián-Jiménez

DOI:
10.55633/s3me/082.2024

Introducción

Los asistentes conversacionales impulsados por inteligencia artificial (IA) han sido diseñados para interactuar con los usuarios de una manera natural. Utilizan el aprendizaje automático y el procesamiento del lenguaje para responder a preguntas que formulan los usuarios sobre cualquier tema, incluyendo los relativos a las ciencias de la salud, pero ninguno de ellos ha sido entrenado, hasta ahora, específicamente en el campo de la toxicología clínica[1]. Modelos como ChatGPT ya han demostrado ser de ayuda en el terreno de la asistencia clínica[2,3], la docencia a los estudiantes de medicina[4], la educación sanitaria de la población[5], la investigación médica[6] y la gestión en el entorno sanitario[7], y son capaces de mejorar la redacción de los artículos científicos que están en fase de preparación[8].

El objetivo principal de este estudio ha sido formular preguntas del ámbito de la toxicología a cuatro sistemas que generan respuestas de texto basadas en IA y a cuatro toxicólogos clínicos (TC) para comprobar si un grupo de observadores, también expertos en toxicología, ha sido capaz de identificar si las respuestas provenían de una IA o de un TC. Esta evaluación de la capacidad de un individuo para discriminar si una respuesta proviene de un humano o de una máquina es el denominado test de Turing[9]. Este se ha propuesto como una prueba que permite catalogar a estos sistemas. Si pasan el test (cuando un elevado porcentaje de sus respuestas son indistinguibles de las respuestas humanas) son considerados como "inteligentes", aunque sea de forma parcial, porque hasta el momento los sistemas disponibles se centran en un aprendizaje automatizado y en razonamientos lógicos en base a parámetros introducidos por operadores humanos. Como objetivos secundarios se valoró la calidad del texto ofrecida por estas IA, así como el nivel de conocimientos que tenían sobre el tema preguntado, en comparación con las respuestas dadas por los TC.

Método

Dos de los autores (SNX y MAT), con experiencia en la asistencia de pacientes intoxicados, elaboraron 10 preguntas de toxicología clínica general que incluían aspectos sobre las manifestaciones clínicas, las repercusiones analíticas y las opciones terapéuticas en casos de intoxicación por medicamentos, drogas de abuso, productos industriales, plantas tóxicas u otras substancias, así como recomendaciones en el campo de la prevención (Tabla 1).

Se seleccionaron cuatro sistemas populares de IA con los que los autores estaban familiarizados por trabajos previos de investigación sobre cómo se desenvuelven estas IA en el terreno de la toxicología (ChatGPT en su versión "3.5", Bard en su versión "1.5", LuzIA en su versión "1.1.0" y Copilot, antiguamente denominada Bing, en su versión "Copilot gratuita"). En todos los casos se trataba de versiones libres que estaban disponibles sin subscripción en enero de 2024. A través de sus respectivas páginas web (https://chat.openai.com/; https://bard.google.com/chat?hl=es; https://web.whatsapp.com/; https://copilot.microsoft.com), se introdujeron las 10 preguntas y se recogieron las respuestas generadas por estos cuatro modelos de lenguaje.

Tabla 1. Preguntas formuladas a los sistemas de inteligencia artificial y a los toxicólogos clínicos

1. ¿Me puedes explicar lo que sucedió en la Bahía de Minamata (Japón) con el mercurio?
2. ¿Me puedes describir las manifestaciones clínicas de una intoxicación aguda por litio?
3. ¿Me puedes decir los signos y síntomas que se producen si se ingieren semillas de *Datura stramonium* y cómo se trata esta intoxicación?
4. ¿Me puedes decir los efectos tóxicos del consumo de LSD?
5. ¿Me puedes describir las intoxicaciones que pueden dar lugar a un electrocardiograma con un patrón de síndrome de Brugada?
6. ¿Me puedes describir las alteraciones analíticas que se pueden observar en una intoxicación por metanol?
7. ¿Me puedes decir en qué se basa la indicación del carbón activado en el tratamiento de las ingestas tóxicas por vía oral?
8. ¿Me puedes explicar lo que es un *body-packer* y sus principales complicaciones?
9. ¿Me puedes resumir lo que es la sensibilidad química múltiple o intolerancia ambiental idiopática?
10. ¿Me puedes decir los 10 puntos clave para prevenir las intoxicaciones de los niños en el hogar?

De forma paralela, se solicitó a cuatro TC españoles (A, B, C y D), que respondiesen a las mismas preguntas. Dado que la especialidad de toxicología clínica no está reconocida en España, los autores decidieron proponer para esta función a cuatro personas que cumplían este triple requisito: 1) ser médico en activo y trabajando como urgenciólogo en el servicio de urgencias de un hospital público español; 2) ser uno de los 10 miembros del patronato de la Fundación Española de Toxicología Clínica o uno de los 18 miembros del comité de expertos que asesoran a la citada Fundación; y 3) tener demostrada una notable actividad asistencial, docente e investigadora en el terreno de las intoxicaciones. Los expertos de la Fundación son nombrados por su patronato en función de la acreditación de sus méritos curriculares en esta rama científica.

En total se dispuso de 40 respuestas de los sistemas de IA a las 10 cuestiones planteadas y de otras tantas respuestas de los TC, lo que permitió hacer comparaciones por pares IA-TC en igualdad de condiciones. Las evaluaciones de la prueba de Turing, así como de la calidad de las respuestas, se realizaron mediante observadores, definidos como personal sanitario con conocimientos notables en toxicología. Para conseguir observadores se solicitó su participación en el estudio a los miembros del Grupo de Trabajo de Toxicología de la Societat Catalana de Medicina d'Urgències i Emergències (SoCMUETox), y a los 24 que aceptaron se les remitió por correo electrónico un cuestionario con las 10 preguntas (Tabla 1) y otros tantos pares de respuestas, de modo que cada pregunta tuviese, de forma aleatoria, una respuesta pro-

Tabla 2. Opciones a seleccionar por parte de los 24 observadores para cada una de las preguntas y respuestas*

Procedencia de la respuestas (escoger solo una de estas opciones)	– Proviene con certeza de una IA – Proviene probablemente de una IA – No se distinguir de quién proviene la respuesta – Proviene probablemente de un TC – Provine con certeza de un TC
Calidad del texto (escoger solo una de estas opciones)	– Texto excelente – Texto con errores u omisiones menores – Texto con errores u omisiones mayores – Texto impresentable
Conocimientos sobre el tema objeto de la pregunta	– Puntuar en un rango entre 0 y 10

*Cada pregunta tenía dos respuestas y siempre había una de un sistema de IA y otra de un TC.
IA: inteligencia artificial; TC: toxicólogo clínico.

cedente de una IA y otra de un TC de forma ciega para el observador. Este tenía que decidir qué respuesta procedía de una IA y qué respuesta era de un TC según el baremo que se muestra en la Tabla 2 para poder evaluar el test de Turing.

Se consideró que un sistema de IA era "inteligente en el campo de la toxicología clínica" si conseguía que por lo menos el 30% de sus respuestas fuesen atribuidas por los observadores, con certeza o con probabilidad a un TC, o cuando afirmaban no saber distinguir la procedencia de la respuesta. Asimismo, los observadores debían valorar la calidad del texto y cuantificar el nivel de conocimientos sobre el tema que se deducía a partir de las respuestas dadas (Tabla 2).

Los resultados se introdujeron en una base de datos SPSS (Versión 26 Armonk, IBM Corp, Nueva York, EE.UU.) para su análisis estadístico por parte de uno de los autores (JRG). Las variables cuantitativas se han descrito mediante la mediana y el rango intercuartil [percentiles 25 y 75], así como el rango absoluto definido como el mínimo y el máximo, excepto en el caso de la valoración de conocimientos que se desprenden de las respuestas, que se empleó la media como tendencia central y la desviación estándar (DE) como medida de variabilidad. En el caso de las variables cualitativas, se expresan como frecuencia absoluta y porcentaje. La comparabilidad entre las IA y los TC en la prueba de Turing se realizó con la variable ordinal original, empleando la prueba de la U de Mann-Whitney y con el resultado cualitativo en relación a si pasaba el test de Turing, definido dicotómicamente como se ha descrito, mediante la prueba exacta de Fisher. La valoración de la calidad de la respuesta se realizó mediante la prueba de la U de Mann-Whitney. Finalmente se comparó la concordancia entre la respuesta de la IA con relación a la ofrecida por TC en dos pasos: se calculó el índice de concordancia Kappa y se comparó la valoración cuantitativa de la calidad emitida por los 24 observadores, estimando las diferencias promedio y el intervalo de confianza al 95% (IC 95%) y evaluando estas diferencias mediante la prueba T de muestras relacionadas. En todos los análisis, se consideró un error de tipo I bilateral del 5%.

Resultados

De los 24 observadores que participaron en el estudio, 19 (79,2%) eran médicos, cuatro (16,7%) enfermeros y un (4,2%) farmacéutico, y procedían de 12 hospitales de diferente complejidad asistencial y también, en cuatro casos, del servicio de emergencias médicas (SEM) de Cataluña.

De las 240 respuestas que analizaron los observadores (10 por observador) y que procedían de alguna IA (Tabla 3), en el 30% de las ocasiones los sistemas de IA "engañaron" a los observadores al no ser identificados como IA. Por tanto, puede afirmarse que pasaron el test de Turing y, por ello, deben considerarse inteligentes en el campo de la toxicología clínica. Dentro de este 30%, en 21 ocasiones (8,8%) los observadores opinaron que con certeza las respuestas provenían de un TC, en 38 (15,8%) que procedían probablemente de un TC y en 13 (5,4%) reconocían que no sabía distinguir si la respuesta era de una IA o de un TC. Este resultado del test de Turing fue heterogéneo entre las IA. Así, LuzIA y ChatGPT mostraron una mayor capacidad de engaño a los observadores, con diferencias estadísticamente significativas respecto a Bard ($p = 0,036$ y $p = 0,041$, respectivamente).

Con relación a la calidad de los textos de las respuestas ofrecidas por las IA (Tabla 3), la valoración global de los observadores fue de "excelente" en el 38,8% de las ocasiones, con una diferencia significativa en favor de ChatGPT (que alcanzó la calificación de excelente en el 61,3% de sus respuestas) respecto a Bard (34,4%, $p < 0,001$), LuzIA (31,7%, $p < 0,001$) y Copilot (26,3%, $p < 0,001$).

Respecto a la percepción de conocimientos sobre el tema que se deducía de las respuestas de las IA, la puntuación media de los cuatro sistemas fue de 7,23 (DE 1,57) sobre 10, obteniendo ChatGPT una puntuación de 8,03 (DE 1,26) que fue significativamente mayor a la obtenida por Luzia (7,02 [DE 1,63], $p < 0,001$), Bard (6,91 [DE 1,64], $p < 0,001$) y Copilot (6,91 [DE 1,46], $p < 0,001$).

De las respuestas que se analizaron y que procedían de un TC (Tabla 4), en el 27,9% de las respuestas los TC "engañaron" a los observadores, ya que estos presentaron serias dudas sobre la procedencia de los textos. En concreto, en 23 (9,6%) ocasiones los observadores consideraron que con certeza provenía de una IA, en 36 (15,0%) que provenía probablemente de una IA y en 8 (3,3%) reconocían que no sabía distinguir si la respuesta provenía de una IA o de un TC. Por tanto, puede afirmarse que los TC no pasaron el test de Turing al no alcanzar el 30%. Pero no todos los TC tuvieron el mismo comportamiento ya que uno de ellos (TC-B) sí que confundió en un 32,8% de las ocasiones a los observadores, aunque no hubo diferencias estadísticamente significativas entre ellos.

Con relación a la calidad del texto de las respuestas proporcionadas por los TC, la valoración global de los observadores fue de excelente en el 57,9% de las ocasiones, con una diferencia estadísticamente significativa

Tabla 3. Valoración de los 24 observadores a las respuestas provenientes de una inteligencia artificial

		Total N = 240 n (%)	Copilot N = 57 n (%)	Bard N = 61 n (%)	LuzIA N = 60 n (%)	ChatGPT N = 62 n (%)	p-valor (todos)	Copilot *vs* Bard	Copilot *vs* LuzIA	Copilot *vs* ChatGPT	Bard *vs* LuzIA	Bard *vs* ChatGPT	LuzIA *vs* ChatGPT
Opinión de los observadores sobre el origen de la respuesta	Proviene con certeza de una IA	61 (25,4)	19 (33,3)	18 (29,5)	12 (20,0)	12 (19,3)	0,102	0,514	0,198	0,176	**0,036**	**0,041**	0,920
	Proviene probablemente de una IA	107 (44,6)	19 (33,3)	32 (52,5)	26 (43,3)	30 (48,4)							
	No sé distinguir	13 (5,4)	5 (8,8)	3 (4,9)	3 (5,0)	2 (3,2)							
	Proviene probablemente de un TC	38 (15,8)	12 (21,0)	4 (6,6)	15 (25,0)	7 (11,3)							
	Proviene con certeza de un TC	21 (8,7)	2 (3,5)	4 (6,6)	4 (6,7)	11 (17,7)							
Test de Turing	Pasa el test de Turing*	72 (30,0)	19 (33,3)	11 (18,0)	22 (36,7)	20 (32,7)	0,117	0,056	0,706	0,901	**0,021**	0,069	0,608
	No pasa el test de Turing*	168 (70,0)	38 (66,7)	50 (82,0)	38 (63,3)	42 (67,7)							
Opinión de los observadores sobre la calidad del texto	Texto excelente	93 (38,7)	15 (26,3)	21 (34,4)	19 (31,7)	38 (61,3)	< **0,001**	0,961	0,434	< **0,001**	0,517	< **0,001**	< **0,001**
	Texto con errores u omisiones menores	101 (42,1)	29 (50,9)	22 (36,1)	30 (50,0)	20 (32,3)							
	Texto con errores u omisiones mayores	38 (15,8)	11 (19,3)	13 (21,3)	10 (16,7)	4 (6,4)							
	Texto impresentable	8 (3,3)	2 (3,5)	5 (8,2)	1 (1,7)	0 (0 ,0)							
Valoración cuantitativa de los observadores sobre los conocimientos de la IA	Mediana (P25-P75)	7,0 [6,0; 8,0]	7,0 [6,0;7,0]	7,0 [6,0;8,0]	7,0 [6,0; 8,0]	8,0 [7,0;9,0]	< **0,001**	0,750	0,487	< **0,001**	0,710	< **0,001**	< **0,001**
	Rango	2,0 a 10,0	2,0 a 10,0	3,0 a 10,0	2,0 a 10,0	5,0 a 10,0							

*La IA pasa el test de Turing si ≥ 30% de los observadores han considerado que la respuesta provenía con certeza o probabilidad de un toxicólogo clínico, o no lo han sabido distinguir.
IA: sistema de inteligencia artificial; TC: toxicólogo clínico; DS: desviación estándar; P25: percentil 25; P75: percentil 75.
Los valores en negrita denotan significación estadística (p < 0,05).

Tabla 4. Valoración de 24 observadores a las respuestas provenientes de un toxicólogo clínico

		Total N = 240 n (%)	Toxicólogo A N = 60 n (%)	Toxicólogo B N = 58 n (%)	Toxicólogo C N = 65 n (%)	Toxicólogo D N = 57 n (%)	p-valor (todos)	A *vs* B	A *vs* C	A *vs* D	B *vs* C	B *vs* D	C *vs* D
Opinión de los observadores sobre el origen de la respuesta	Proviene con certeza de una IA	23 (9,6)	7 (11,7)	3 (5,2)	7 (0,8)	6 (10,5)	0,773	0,333	0,415	0,643	0,836	0,616	0,740
	Proviene probablemente de una IA	36 (15,0)	6 (10,0)	13 (22,4)	9 (13,8)	8 (14,0)							
	No sé distinguir	8 (3,3)	2 (3,3)	3 (5,2)	2 (3,1)	1 (1,7)							
	Proviene probablemente de un TC	107 (44,6)	25 (41,7)	25 (43,1)	31 (47,7)	26 (45,6)							
	Proviene con certeza de un TC	66 (27,5)	20 (33,3)	14 (24,1)	16 (24,6)	16 (28,1)							
Test de Turing	Pasa el test de Turing*	67 (27,92	15 (25,0)	19 (32,8)	18 (27,7)	15 (26,3)	0,80	0,352	0,733	0,871	0,541	0,449	0,864
	No pasa el test de Turing*	173 (72,1)	45 (75,0)	39 (67,2)	47 (72,3)	42 (73,7)							
Opinión de los observadores sobre la calidad del texto	Texto excelente	139 (57,9)	38 (63,3)	37 (63,8)	37 (56,9)	27 (47,4)	0,178	0,919	0,348	**0,047**	0,469	0,073	0,246
	Texto con errores u omisiones menores	83 (34,6)	20 (33,3)	17 (29,3)	23 (35,4)	23 (40,3)							
	Texto con errores u omisiones mayores	16 (6,7)	2 (3,3)	3 (5,2)	5 (7,7)	6 (10,5)							
	Texto impresentable	2 (0,8)	0 (0,0)	1 (1,7)	0 (0,0)	1 (1,7)							
Valoración cuantitativa de los observadores sobre el nivel de conocimientos de la IA sobre el tema objeto de la pregunta	Mediana (P25-P75)	8,0 [7,0;9,0]	8,7 [7,0;9,0]	8,0 [7,0;9,0]	8,0 [7,0; 9,0]	8,0 [7,0;9,0]	0,205	0,661	0,082	0,088	0,218	0,224	0,923
	Rango	2,0 a 10,0	6,0 a 10,0	5,0 a 10,0	5,0 a 10,0	2,0 a 10,0							

*El TC pasa el test de Turing si ≥ 30% de los observadores han considerado que la respuesta provenía con certeza o probabilidad de una IA o no lo han sabido distinguir.
IA: sistema de inteligencia artificial; TC: toxicólogo clínico; ; DS: desviación estándar; P25: percentil 25; P75: percentil 75.
Los valores en negrita denotan significación estadística (p < 0,05).

entre el TC-A (63,3% de textos excelentes) y el TC-D (47,4% de excelencias, p = 0,047).

Respecto a la percepción de conocimientos sobre el tema que se deducía de las respuestas de los TC, la puntuación media fue de 7,9 (1,4) sobre un máximo de 10, obteniendo el TC-A la mejor valoración de 8,2 (1,2) pero sin diferencias significativas respecto a sus colegas.

Finalmente se comparó la calidad de los textos y el nivel de conocimientos entre las IA y los TC (Tabla 5). Con relación a la calidad de los textos percibida por los 24 evaluadores, no se ha observado concordancia entre las respuestas provenientes de una IA y las de un TC (índice Kappa 0,011), ya que los observadores consideraron que 101 (42,1%) de las 240 respuestas analizadas eran mejores las ofrecidas por los TC, 46 (19,2%) respuestas eran mejores si provenían de una IA y en 93 (38,8%) no observaron diferencias. El análisis individualizado de las 10 respuestas muestra la misma falta de concordancia.

Con relación al nivel de conocimientos, la puntuación fue buena en ambos grupos y superó el promedio de 7, pero hubo una diferencia estadísticamente significativa de forma global en favor de los TC ($p < 0,001$) y en particular en las preguntas 2, 3, 6, 7 y 8. Solo la preguntas 4 y 9 obtuvieron mejor puntuación (no significativa) de conocimiento cuando la respuesta provenía de una IA.

Discusión

Desde que John McCarthy acuñase en 1956 el término "inteligencia artificial" para describir a las máquinas que son capaces de realizar tareas que normalmente se atribuyen al hombre, como resolver problemas, la comunidad científica ha ido en busca de alguna prueba que permita etiquetar a estos sistemas como inteligentes[10]. La solución más aceptada, aunque discutida, a este dilema es el denominado test de Turing en honor a Alan Turing, un matemático inglés que en los años 50 del pasado siglo propuso una prueba basada en el juego de la imitación con el objetivo de determinar si una máquina puede comportarse de modo similar a un humano[11]. Para realizar esta prueba, la máquina y un humano generan textos por escrito en respuesta a una serie de preguntas y si un evaluador externo es incapaz de distinguir, en un número considerable de ocasiones, si el texto procede de un humano o de la máquina, se considera que esta pasa el test y puede considerarse inteligente. No existe un acuerdo unánime respecto al porcentaje de respuestas que marca el punto de inflexión en esta prueba, pero a efectos del presente estudio se estableció como punto de corte el 30%, como ya han hecho otros autores[12,13].

La mayoría de los sistemas de IA disponibles a fecha de hoy (julio 2024) no han sido entrenados específicamente en el campo de la toxicología. Ello les podría hacer más susceptibles a ser identificados como máquinas cuando, como ocurre en el presente estudio, los observadores son expertos en esta materia. A pesar de ello, tres de las cuatro IA han pasado el test de Turing (≥ 30% de sus respuestas han sido interpretadas como provenientes de un TC, o los observadores no lo han sabido distinguir), lo que demuestra su alta capacidad para recoger información sobre temas de toxicología y, sobre todo, para usar un lenguaje muy natural y bien estructurado al exponer las respuestas. Destaca entre las cuatro LuzIA, que ha sido la IA con mayor apariencia de TC en sus respuestas.

Estudios previos realizados por nuestro grupo ya habían demostrado que estos cuatro sistemas de IA tienen capacidad para aprobar, y algunos con buena nota, un examen de toxicología clínica como el que se pone a los estudiantes de medicina[14]. Asimismo, verificamos que estos sistemas son capaces de realizar diagnósticos toxicológicos precisos cuando se les presenta un caso clínico publicado en la literatura médica[15]. Dado que la respuesta de estos sistemas está guiada por un conocimiento previo disponible en internet, estos resultados pueden estar condicionados, no por el aprendizaje, sino por la repetición de textos escritos en el pasado. Por ello, ha de quedar claro que, aun pasando el test de Turing o la prueba de la señal de inteligencia mínima[16] o la prueba de Ebert[17], ninguno de estos sistemas es realmente inteligente. Aunque sean capaces de conversar con un humano de una manera casi indistinguible de otro humano por su avanzada tecnología, consistente en el pronóstico de las siguientes palabras que debe poner en un texto gracias a la enorme cantidad de información puesta a su disposición. Pero, a pesar de ello, son una herramienta potencial útil en muchos campos, incluida la medicina en general[18,19] y la toxicología clínica en particular[20,21].

La mayoría de los observadores manifestaron en sus comentarios las dificultades que había tenido para identificar la procedencia de los textos y, con frecuencia, no era la calidad de la respuesta sino pequeños detalles en la jerga que los médicos utilizan los que hacían decantar al observador hacia que el texto provenía de un TC. Además, cuanto más se use una terminología específica de una especialidad como la toxicología, más probable es que el texto esté escrito por un TC[22], mientras que si se detectan errores de sentido común, de relevancia, de razonamiento y de lógica, más probable que el texto provenga de una IA[23].

En general, la calidad de los textos generados por las IA ha sido catalogada de buena o muy buena (38,8%), aunque los TC han superado esta cifra (57,9%). Los observadores han considerado que el nivel de conocimientos de las IA sobre el tema objeto de la pregunta era alto (7,23 sobre 10), aunque también se han visto superadas por los TC (7,94), aunque con una magnitud marginal en términos cualitativos. Por tanto, aunque las IA se sabe que cometen errores al dar sus respuestas, ofrecen un nivel de fiabilidad bastante aceptable en el terreno de la toxicología clínica, ya que solo un 15,8% de los observadores han considerado que sus textos contenían errores u omisiones mayores y un 3,3% los han calificado como impresentables. Esto tam-

Tabla 5. Comparación de las 10 respuestas provenientes de una IA y de un TC en relación a la calidad del texto y al nivel de conocimientos

Pregunta			TC Calidad				Total	Índice Kappa de concordancia	Valoración cuantitativa		Diferencias (CI 95%)	Valor de p
			Excelente	Error u omisión menor	Error u omisión mayor	Impresentable			IA Conocimientos	TC Conocimientos		
1	IA Calidad	Excelente	7	5	0	0	12	0	7,8	8,1	–0,4 (–1,1; 0,3)	0,273
		Error u omisión menor	7	3	0	0	10					
		Error u omisión mayor	0	2	0	0	2					
	Total		14	10	0	0	24					
2	IA Calidad	Excelente	3	2	1	0	6	0,007	6,5	7,8	–1,3 (–2,2; –0,3)	**0,009**
		Error u omisión menor	2	4	1	0	7					
		Error u omisión mayor	6	5	0	0	11					
	Total		11	11	2	0	24					
3	IA Calidad	Excelente	5	1	0	0	6	0,045	6,8	8,4	–1,6 (–2,6; –0,6)	**0,004**
		Error u omisión menor	9	3	0	0	12					
		Error u omisión mayor	2	1	0	0	3					
		Impresentable	3	0	0	0	3					
	Total		19	5	0	0	24					
4	IA Calidad	Excelente	4	7	1	0	12	0,071	7,9	7,3	0,5 (–0,2; 1,3)	0,168
		Error u omisión menor	2	7	1	0	10					
		Error u omisión mayor	2	0	0	0	2					
	Total		8	14	2	0	24					
5	IA Calidad	Excelente	5	2	2	0	9	0,008	7,2	7,6	–0,4 (–1,2; 0,4)	0,336
		Error u omisión menor	6	4	2	0	12					
		Error u omisión mayor	2	1	0	0	3					
	Total		13	7	4	0	24					
6	IA Calidad	Excelente	5	1	2	0	8	0,166	6,2	7,7	–1,5 (–2,7; –0,3)	**0,018**
		Error u omisión menor	3	4	0	0	7					
		Error u omisión mayor	2	2	1	0	5					
		Impresentable	3	0	1	0	4					
	Total		13	7	4	0	24					
7	IA Calidad	Excelente	6	1	1	0	8	0,013	7,0	8,6	–1,6 (–2,4; –0,8)	< **0,001**
		Error u omisión menor	5	2	0	0	7					
		Error u omisión mayor	6	2	0	0	8					
		Impresentable	1	0	0	0	1					
	Total		18	5	1	0	24					
8	IA Calidad	Excelente	4	3	0	0	7	0	7,4	8,3	–0,9 (–1,5; –0,3)	0,005
		Error u omisión menor	10	7	0	0	17					
	Total		14	10	0	0	24					
9	IA Calidad	Excelente	6	5	0	1	12	0	7,6	7,3	0,3 (–0,7; 1,3)	0,529
		Error u omisión menor	5	4	2	0	11					
		Error u omisión mayor	1	0	0	0	1					
	Total		12	9	2	1	24					
10	IA Calidad	Excelente	8	3	1	1	13	0	7,8	8,2	–0,4 (–1,3; 0,6)	0,413
		Error u omisión menor	7	1	0	0	8					
		Error u omisión mayor	2	1	0	0	3					
	Total		17	5	1	1	24					
Total	IA Calidad	Excelente	53	30	8	2	93	0,011	7,23	7,94	–0,71 (–0,99; –0,43)	< **0,001**
		Error u omisión menor	56	39	6	0	101					
		Error u omisión mayor	23	14	1	0	38					
		Impresentable	7	0	1	0	8					
	Total		139	83	16	2	240					

IA: inteligencia artificial; TC: toxicólogo clínico; IC: intervalo de confianza.
Los valores en negrita denotan significación estadística ($p < 0,05$).

bién ha pasado en los textos de los TC, algunos de los cuales han merecido estos calificativos, aunque en menor proporción (6,7% y 0,8%, respectivamente).

Si se analiza el nivel de concordancia en la calidad percibida por los evaluadores entre cada una de las respuestas de la IA y de los TC, se observa que en la mayoría de ellas es muy bajo (en todas las preguntas es < 0,2 y en cuatro casos de 0). Por tanto, hay una nítida diferencia de valoración, que se ha resuelto siempre en favor de los TC. Es de esperar que la futura e inminente mejora de estos sistemas de IA hará reducir estas diferencias y aumentar por tanto el nivel de concordancia.

Por otro lado, el nivel de conocimientos que transmiten en sus respuestas tanto las IA como los TC es satisfactorio, oscilando del aprobado (nota mínima 6,2) al notable (nota máxima 7,9) en las respuestas de las IA, mientras los TC obtenían notas entre el notable (nota mínima 7,3) y el sobresaliente (nota máxima 8,6). No hubo suspensos (notas < 5) en ninguno de los dos grupos.

Por todo ello, los textos de toxicología creados por las IA podrían ser un punto de partida en algunos aspectos de un documento asistencial, docente o investigador, pero es inexcusable su revisión para estar seguros de que, por un lado, se ajustan al conocimiento científico real y actual y, por otro y aún más importante, no contiene falsedades o inexactitudes[24]. De este modo, ChatGPT podría participar no solo en la redacción de un documento asistencial describiendo, por ejemplo, las características de la enfermedad del paciente[25]. También ha demostrado su competencia para diagnosticar correctamente a enfermos con problemas médicos que acuden a un servicio de urgencias[26]. A nivel docente, cualquiera de estos sistemas de IA podría preparar preguntas de examen o evaluar la calidad de las preguntas propuestas por el profesor. Es conocida la falta de profesores universitarios en el ámbito de las ciencias de la salud[27] y las IA van a ser, sin duda, una herramienta de ayuda y con aplicación inmediata para docentes y alumnos. En temas de investigación, estos sistemas pueden proponer una metodología a partir de los objetivos del estudio, pueden realizar búsquedas bibliográficas muy perfiladas y pueden contribuir a discutir los resultados obtenidos[28]. Creemos también que la ayuda que ya presta la IA en el campo de las ciencias de la salud en general y de la toxicología clínica en particular complementa, pero no sustituye, a la consulta sobre la información fiable y muy actualizada y que se encuentra disponible ya sea en bases de datos generales de alta calidad científica como pueden ser PubMed (https://pubmed.ncbi.nlm.nih.gov), UpToDate (https://sso.uptodate.com) o la Cochrane Library (https://www.cochranelibrary.com), o en bases de datos específicas para toxicología clínica como Poisindex (https://www.micromedexsolutions.com) o ToxBase (https://www.toxbase.org).

El estudio tiene varias limitaciones. En primer lugar, se ha valorado la calidad de las respuestas sobre una materia científica muy concreta, la toxicología clínica y, por tanto, nuestros resultados no son extrapolables a otros campos de la medicina de urgencias y tampoco a otras especialidades[29]. En segundo lugar, solo se han tenido en cuenta las respuestas de cuatro IA y cuatro TC a 10 preguntas, por lo que si el número de valoraciones o de observadores hubiese sido mayor quizás se hubiese alcanzado significación estadística en más variables. En tercer lugar, los observadores son miembros de un grupo de trabajo toxicológico y esto introduce un sesgo en la valoración de las respuestas. Y finalmente, el test de Turing tiene sus propias limitaciones, ya que valora la capacidad de una máquina para imitar respuestas humanas, pero no puede juzgar la creatividad, la conciencia o la comprensión emocional que son componentes de la inteligencia humana[30].

En conclusión, Luzia, Copilot y ChatGPT son sistemas de IA que pasan el test de Turing y pueden considerarse inteligentes en el campo de la toxicología al ser capaces, en más del 30% de las ocasiones, de generar respuestas que parecen haber sido realizadas por un TC. La calidad de los textos generados por estas IA es buena y muy buena en el caso de ChatGPT. El nivel de conocimientos de estos sistemas artificiales en el terreno de la toxicología es alto, pero en general es inferior al de los TC, ya que solo ChatGPT consiguió superar a dos de los TC.

Conflicto de intereses: Los autores declaran no tener conflicto de interés en relación con el presente artículo.

Financiación: Los autores declaran la no existencia de financiación en relación con el presente artículo.

Responsabilidades éticas: Todos los autores han confirmado el mantenimiento de la confidencialidad y respeto de los derechos de los pacientes en el documento de responsabilidades del autor, acuerdo de publicación y cesión de derechos a EMERGENCIAS.

Artículo no encargado por el Comité Editorial y con revisión externa por pares.

Adenda

Investigadores del Grupo SoCMUETox que han contribuído a este estudio: África de la Cruz-Ramos, Albert Moreno-Destruels, Daniel Martínez-Millán y Vicenç Ferrés-Padrós (SEM de la Generalitat de Catalunya); Carlos García-Gutiérrez, Concepción Moll-Tuduri, Emilio Salgado-García, Miguel Galicia-Paredes y Ona Escoda-Turón (Hospital Clínic de Barcelona); Alma M Palomino-Bustos, August Supervía-Caparrós, M Jesús López-Casanova y Rosana Muñoz-Bermúdez (Hospital del Mar, Barcelona); Cristina Ramió-Lluch, M Àngels Gispert-Ametller y Raquel Aguilar-Salmerón (Hospital Dr. Josep Trueta, Girona); Héctor Hernández-Ontiveros e Indalecio Morán-Chorro (Hospital de la Santa Creu i Sant Pau, Barcelona); Josep Piqueras-Carrasco y Lluis Marruecos-Sant (SocMUETox); Lidia Martínez-Sánchez (Hospital de Sant Joan de Deu (Esplugues de Llobregat, Barcelona); Susana Vert-García (Hospital de Viladecans, Barcelona); Francisca Córdoba-Ruíz (Hospital Dr. Moisés Broggi, Sant Joan Despí, Barcelona); Lidia García-Gibert (Hospital Parc Taulí, Sabadell, Barcelona); Ana Rodríguez-Ruitiña (Hospital de Sant Joan de Deu, Manresa, Barcelona); Jordi Puiguriguer Ferrando (Hospital Son Espases, Palma de Mallorca) y Elena Fuentes-González (Hospital de Bellvitge, Hospitalet de Llobregat, Barcelona).

Bibliografía

1 Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell. 2023;6:1169595.
2 Fernández-Topham J, Hernández-Tejedor A, Sánchez-Blasco D, Corral-Torres E. Predicción mediante inteligencia artificial del pronós-

tico neurológico del paciente durante la parada cardiaca extrahospitalaria. Emergencias. 2024;36:233-4.
3 Liu Z, Zhang L, Wu Z, Yu X, Cao C, Dai H, et al. Surviving ChatGPT in healthcare. Front Radiol. 2024;3:1224682.
4 Mohammad B, Supti T, Alzubaidi M, Shah H, Alam T, Shah Z et al. The pros and cons of using ChatGPT in medical education: A scoping review. Stud Health Technol Inform. 2023;305:644-7.
5 Sallam M. ChatGPT utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns. Healthcare (Basel). 2023;11:887.
6 Ruksakulpiwat S, Kumar A, Ajibade A. Using ChatGPT in medical research: Current status and future directions. J Multidiscip Healthc. 2023;16:1513-20.
7 Wang X, Sanders HM, Liu Y, Seang K, Tran BX, Atanasov AG et al. ChatGPT: promise and challenges for deployment in low- and middle-income countries. Lancet Reg Health West Pac. 2023;41:100905.
8 Huang J, Tan M. The role of ChatGPT in scientific communication: writing better scientific review articles. Am J Cancer Res. 2023;13:1148-54.
9 Mintz Y, Brodie R. Introduction to artificial intelligence in medicine. Minim Invasive Ther Allied Technol. 2019;28:73-81.
10 Muthukrishnan N, Maleki F, Ovens K, Reinhold C, Forghani B, Forghani R. Brief history of artificial intelligence. Neuroimaging Clin N Am. 2020;30:393-9.
11 Turing AM. Computing machinery and intelligence. Mind. 1950;59:433-60.
12 Mann A. That computer actually got an F on the Turing Test. (Consultado 7 Mayo 2024). Disponible en: https://www.wired.com/2014/06/turing-test-not-so-fast/
13 Feigenbaum EA. Some challenges and grand challenges for computational intelligence. J ACM. 2003;50:32-40.
14 Nogué-Xarau S, Amigó-Tadín M, Ríos-Guillermo J. Evaluación de los conocimientos de varios sistemas de inteligencia artificial sobre una subespecialidad de la medicina de urgencias y emergencias: la toxicología clínica. Rev Esp Urg Emerg. 2024;3:15-9.
15 Nogué-Xarau S, Amigó-Tadín M, Ríos-Guillermo J. ¿Puede la inteligencia artificial ayudar al urgenciólogo en el diagnóstico de las intoxicaciones?. Emergencias. 2024;36:153-6.
16 McKinstry C. Minimum intelligent signal test: An alternative Turing test. Can Artif Intell. 1997;41:1.
17 Wikipedia. Prueba de Turing. (Consultado 8 Mayo 2024). Disponible en: https://es.wikipedia.org/wiki/Prueba_de_Turing#cite_note-twsL37-98
18 Sedaghat S. Early applications of ChatGPT in medical practice, education and research. Clin Med (Lond). 2023;23:278-9.
19 Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell. 2023;6:1169595.
20 Sabry Abdel-Messih M, Kamel Boulos MN. ChatGPT in Clinical Toxicology. JMIR Med Educ. 2023;9:e46876.
21 Altamimi I, Altamimi A, Alhumimidi AS, Altamimi A, Temsah MH. Snakebite advice and counseling from artificial intelligence: An acute venomous snakebite consultation With ChatGPT. Cureus. 2023;15:e40351
22 Reddy R. Texto generado por IA frente a texto escrito por humanos: análisis completo. (Consultado 23 Mayo 2024). Disponible en: https://www.ranktracker.com/es/blog/ai-generated-vs-human-written-text-complete-analysis/
23 Fischler D. Real or fake text? We can learn to spot the difference. Disponible en: https://techxplore.com/news/2023-02-real-fake-text-difference.html. (Consultado el 23 mayo 2024).
24 Szolovits P. Large language models seem miraculous, but science abhors miracles. N Eng J Med AI. 2024;1(6). doi: 10.1056/Alp230010.
25 Kim HW, Shin DH, Kim J, Lee GH, Cho JW. Assessing the performance of ChatGPT's responses to questions related to epilepsy: A cross-sectional study on natural language processing and medical information retrieval. Seizure. 2024;114:1-8.
26 Hoppe JM, Auer MK, Strüven A, Massberg S, Stremmel C. ChatGPT with GPT-4 outperforms Emergency Department physicians in diagnostic accuracy: Retrospective analysis. J Med Internet Res. 2024;26:e56110.
27 Castro-Delgado R, Burillo-Putze G. Medicina de Urgencias y Emergencias y Universidad. Emergencias. 2024;36:67-9.
28 Calero Sánchez M, González González JC, Sánchez Berriel I, Burillo-Putze G, Roda García JL. El procesamiento de lenguaje natural en la revisión de literatura científica. Rev Esp Urg Emerg. 2024;3:184-95.
29 Castro-Delgado R, Pardo Ríos M. La inteligencia artificial y los servicios de urgencias y emergencias: debemos dar un paso adelante. Emergencias. 2024;36:145-7.
30 French RM. The Turing test: the first 50 years. Trends Cogn Sci. 2000;4:115-22.