REUE | Original Article

Assessing 4 artificial intelligence systems' knowledge of a subspecialty of emergency medicine: clinical toxicology

Santiago Nogué-Xarau[1], Montserrat Amigó-Tadín[2], José Ríos-Guillermo[3]

BACKGROUND AND OBJECTIVE. Artificial intelligence (AI) is a branch of computer technology that develops systems able to perform tasks associated with human intelligence. The main objective of this study was to evaluate AI answers to questions related to clinical toxicology.

MATERIALS AND METHODS. We evaluated 4 AI applications: ChatGPT, Bing, LuzIA, and Bard. Thirty multiple-choice test questions in Spanish about various aspects of clinical toxicology were presented to the applications, and the answers were assessed. Each question included 5 possible answers, 1 of which was correct. In addition to correctness, we evaluated the bibliographic support each application provided. If the application gave an incorrect answer, we rephrased the question, presented it again, and reevaluated the new answer to detect whether question quality influenced performance. Data were recorded for analysis with SPSS. The level of statistical significance was set at $P < .05$.

RESULTS. The scores achieved by the AI applications were as follows: Bing, 70%; ChatGPT and LuzIA, 67% each; and Bard, 57% ($P > .05$). The scores improved after the incorrect questions were rephrased, but the differences were not significant. Bing included direct access to 3 references per question and Bard to 4. However, only 7.2% and 0.85% of the references, respectively, were to PubMed-indexed sources.

CONCLUSIONS. All 4 AI applications were able to correctly answer more than half the questions about clinical toxicology. After rephrasing some questions, each system achieved more correct answers. The supporting references the applications provided were few and of poor quality.

Keywords: Artificial intelligence. Toxicology. Education. Test questions.

Evaluación de los conocimientos de varios sistemas de inteligencia artificial sobre una subespecialidad de la medicina de urgencias y emergencias: la toxicología clínica

OBJETIVO. La inteligencia artificial (IA) es una disciplina de la informática que se encarga de crear sistemas capaces de realizar tareas que se atribuyen a la inteligencia humana. El objetivo principal de este estudio ha sido evaluar las respuestas de algunas IA a preguntas del campo de la toxicología clínica (TC).

MATERIAL Y MÉTODOS. Se han valorado cuatro aplicaciones de IA: ChatGPT, Bing, LuzIA y Bard. Para evaluar sus conocimientos en TC se les formularon 30 preguntas sobre diversos aspectos de la TC. Cada pregunta ofrecía cinco opciones de respuesta, de las cuales sólo una era correcta. Se evaluó el acierto/error en la respuesta, así como si había apoyo bibliográfico. Si se detectaban respuestas erróneas, se reformuló la misma pregunta, pero utilizando otra forma de lenguaje para evaluar de nuevo la respuesta y ver si la misma era sensible a la calidad de la pregunta. Los datos se introdujeron en una base SPSS para su análisis estadístico. Se consideró significativo un valor de $p < 0,05$.

RESULTADOS. Los porcentajes de respuestas acertadas fueron del 70% (Bing), 67% (ChatGPT y LuzIA) y 57% (Bard), sin diferencias estadísticamente significativas. Al reformular las preguntas en los casos en los que la respuesta de la IA había sido errónea, los porcentajes de aciertos subieron en los cuatro sistemas, pero sin diferencias significativas. En sus respuestas, Bing ofreció el acceso directo a tres citas bibliográficas y Bard a cuatro, pero su presencia en PubMed era muy baja (7,2% y 0,85% respectivamente).

CONCLUSIONES. Los cuatro sistemas de IA han mostrado una capacidad de acierto en más del 50% de las preguntas formuladas de TC. No obstante, el soporte bibliográfico que proporcionan es escaso y de muy baja calidad.

Palabras clave: Inteligencia artificial. Toxicología. Docencia. Preguntas de examen.

Author Affiliations: [1]Fundación Española de Toxicología Clínica, Barcelona, Spain. [2]Área de Urgencias, Hospital Clínic, Barcelona, Spain. [3]Departamento de Farmacología Clínica, Hospital Clínic, Barcelona, Spain.
Corresponding Author: Santiago Nogué-Xarau. Fundación Española de Toxicología Clínica.
E-mail: snoguex@gmail.com
Article Information: Received: 21-12-2023. Accepted: 26-12-2023. *Online:* 3-1-2024.
Editor in Charge: Guillermo Burillo-Putze.

Introduction

Artificial intelligence (AI) is a new field within computer science characterized by its ability to perform tasks that simulate human intelligence, including the rapid resolution of complex problems.[1] These systems require large volumes of data for training, as well as mathematical and statistical algorithms and models that enable them to process this information. All of this operates within supercomputers equipped with high data storage and processing capacity, allowing them to perform trillions of operations per second.[2]

The uses of AI systems in everyday life are numerous and include web-based mapping applications that indicate the best route between locations, virtual assistants (chatbots), spam filters in email, facial recognition in mobile devices, and automatic language translation systems, among others.[3]

In university education, AI could potentially provide individualized tutoring for students through chatbots capable of answering questions, offering guidance on available resources, or suggesting learning strategies.[4] AI can also assist in grading exams by identifying plagiarism or highlighting errors that a human grader might overlook due to factors such as fatigue.

In the field of health sciences, AI could help identify risk factors for disease within specific populations by analyzing their health data. It may also assist in reviewing imaging modalities such as mammograms to support radiologists in screening for suspicious cancer-related findings and could enhance patient care through electronic devices capable of detecting cardiac arrhythmias,[5,6] among other applications. In emergency and critical care medicine, AI-based tools have been evaluated to aid in suggesting diagnoses for the most prevalent diseases encountered in emergency departments. Although early results are promising, such systems still require clinical supervision by the attending medical team.[7]

One subspecialty within emergency medicine is clinical toxicology (CT),[8] a discipline focused on the diagnosis and treatment of poisonings. The most widely available AI systems have not been specifically trained in CT; therefore, the primary objective of this study was to evaluate the ability of several AI systems to answer questions related to clinical toxicology, and as secondary objectives, to determine whether these systems are sensitive to the quality of the question posed, to analyze the speed and depth of their responses, and to assess the bibliographic support they provide.

Material and methods

We evaluated a total of 4 AI systems based on 3 inclusion criteria: free access, real-time question–answer interaction through a chat interface, and operation in the Spanish language. Among the available options, ChatGPT, Bing, LuzIA, and Bard were selected, as the authors were most familiar with these applications (Table 1). All evaluations were performed during May 2023.

To assess the adequacy of the responses in relation to clinical toxicology, the systems were tested with a multiple-choice examination consisting of 30 questions, identical to the one administered to Toxicology students at the School of Medicine, *Universidad de Barcelona* (Barcelona, Spain), in June 2019. Each question offered five possible answers, only one of which was correct. The questions and answer options were prepared by one of the authors (SN). Each question was entered into the AI chat, requesting the system to indicate which of the five options it considered correct. The AI responses—provided in text format—were categorized as correct or incorrect by consensus between two expert authors in CT (SN and MA).

The response latency (in seconds) and length of response (in number of words) were recorded. It was also noted whether the responses included bibliographic references, and in such cases, whether the citations were real or fabricated and, if real, whether they were indexed in PubMed.

If incorrect responses were detected, the corresponding question was reviewed and rephrased to improve clarity and remove possible ambiguities. The reformulated question, with the same answer options, was resubmitted to the AI system that had failed, to assess whether this modification altered its response.

Statistical analysis (performed by author JR) was conducted using SPSS software (version 26). Results were expressed as absolute numbers, percentages (%), or medians with interquartile range (IQR; 25th–75th percentile) for qualitative or quantitative variables, respectively. The correctness (correct/incorrect) of responses was evaluated using the McNemar test. To compare response times and word counts, a Generalized Estimating Equation (GEE) model was applied, accounting for intra-AI variability (as each question was answered by all four systems). Quantitative variables were converted to ranks for a nonparametric approximation. A P value < .05 was considered statistically significant.

Table 1. Methodological aspects of the study: evaluated artificial intelligence systems

Name	Version	Developer	Electronic access link
ChatGPT	3.5	OpenAI	https://chat.openai.com Last accessed December 23rd, 2023
Bing	Version available in May 2023	Microsoft	https://www.bing.com Last accessed December 23rd, 2023
LuzIA	Version available in May 2023	Fundación LuzIA	https://web.whatsapp.com Last accessed December 23rd, 2023
Bard	1.0	Google	https://bard.google.com/chat?hl=es Last accessed December 23rd, 2023

Table 2. Results obtained from artificial intelligence responses (n = 30) to clinical toxicology questions (n = 4)

			ChatGPT	Bing	Luzia	Bard
Accuracy–error when answering	Correct	N (%)	17 (56.67)	21 (70.00)	20 (66.67)	20 (66.67)
	Incorrect	N (%)	13 (43.33)	9 (30.00)	10 (33.33)	10 (33.33)
		P value vs ChatGPT		1.000	1.000	0.581
		P value vs Bing			1.000	0.344
		P value vs LuzIA				0.607
Response latency (seconds)		Median [P25th; P75th]	5.78 [4.42; 6.73]	9.81 [8.00; 10.90]	21.54 [18.11; 25.61]	12.03 [7.08; 15.53]
		Range	3.56 to 10.17	5.43 to 21.08	13.79 to 36.13	4.58 to 19.64
		n	30	30	30	30
		P value vs ChatGPT		< 0.001	< 0.001	< 0.001
		P value vs Bing			0.248	< 0.001
		P value vs LuzIA				< 0.001
Number of words per response		Median [P25th; P75th]	139 [110; 168]	21 [14; 35]	139 [95; 171]	15 [12; 30]
		Range	42 to 213	6 to 96	11 to 199	6 a 65
		n	30	30	30	30
		P value vs ChatGPT		< 0.001	< 0.001	0.672
		P value vs Bing			0.128	< 0.001
		P value vs LuzIA				< 0.001
Review and reformulation of incorrectly answered questions: new accuracy–error rate	Correct	N (%)	19 (63.33)	24 (80.00)	23 (76.67)	25 (83.33)
	Incorrect	N (%)	11 (36.67)	6 (20.00)	7 (23.33)	5 (16.67)
		P value vs ChatGPT		1.000	0.625	0.289
		P value vs Bing			1.000	0.125
		P value vs LuzIA				0.109
		P value comparing each AI before and after reformulated questions	0.250	0.250	0.063	0.500
No. of bibliographic references per response (n [IQR])			0	3 [2.75; 4.00]	0	4 [3.00; 4.25]
Total No. of references across 30 responses (n)			0	97	0	118
Verified existing references, n (%)			Not applicable	97 (100)	Not applicable	12 (10.17)
Real references from medical, pharmacy, or nursing journals, n (% of total)			Not applicable	24 (24.7)	Not applicable	1 (0.85)
Real references indexed in PubMed (% of total)			Not applicable	7 (7.2%)	Not applicable	1 (0.85%)

IQR: interquartile range (25th–25th percentiles).

Results

The main results are presented in Table 2. All 4 AI systems passed the toxicology test, with accuracy rates ranging from 56.7% (ChatGPT) to 70% (Bing), with no statistically significant differences among them. Medical students achieved 72% correct answers on the same examination. All four AIs answered 17 questions (56.7%) correctly, 5 questions (16.7%) incorrectly, and in the remaining 8 questions (26.7%), some systems responded correctly while others failed.

Response initiation times were very fast, particularly for ChatGPT (median 5.78 seconds), which was significantly faster than the other systems ($P < .001$). Regarding verbosity, ChatGPT and LuzIA produced notably longer answers, whereas Bing and Bard were much more concise ($P < .001$).

When incorrectly answered questions were reformulated using clearer and more precise language, all four systems improved their performance.

Neither ChatGPT nor LuzIA provided bibliographic references to support their answers. Bing included a median of 3 references per response; all were real, and 24.7% corresponded to journals in health sciences (medicine, nursing, or pharmacy), though only 7.2% were indexed in PubMed. Bard, in contrast, supplied a median of 4 references per response, but 89.8% were fabricated, as neither the title, authors, nor journal matched actual publications, and only 0.85% were indexed in PubMed.

Discussion

All 4 AI systems demonstrated substantial knowledge of clinical toxicology, successfully passing the proposed examination. Medical students at the *Universidad de Barcelona* achieved a 72% correct response rate on the same test. Other authors have previously shown that these systems can even pass the United States Medical Licensing Examination (USMLE), as ChatGPT-4 answered correctly in 90% of the questions.[9] This level of performance suggests a potential for AI systems to assist in professional medical practice, particularly when managing databases containing thousands of patients[10] or when dealing with complex clinical cases.[11] Their potential has already been highlighted in several medical fields, including forensic medicine,[12] public health,[13] pharmacology,[14] ophthalmology,[15] cardiology,[16] oncology,[17] pulmonology,[18] neurology,[19] neurosurgery,[20] hepatology,[21] dentistry,[22] and angiology,[23] among others.[24] Nonetheless, these AI systems—despite their sophisticated natural language capabilities based on vast datasets—are not infallible. They make mistakes resulting from a lack of genuine comprehension of the text they generate, which could lead to malpractice in clinical settings.[25] The World Health Organization (WHO) has published a report on the ethics of AI in health, outlining 6 guiding principles for its design and use: to protect human autonomy; promote human well-being; ensure transparency; foster accountability; guarantee fair-

ness; and enhance safety.[26] Nonetheless, compliance with these principles is uncertain, as the proprietary algorithms underlying these AI systems are not publicly disclosed, preventing verification of their data sources.

In university education, AI could benefit both educators and students.[27,28] Regarding student assessment, our study shows that refining the phrasing of questions improved AI accuracy—indicating that question quality directly influences response accuracy, a fact already well known among educators. It might therefore be useful for instructors to pre-test their exams through an AI system to identify questions that yield incorrect responses, as such questions may be ambiguously formulated and could be improved to eliminate interpretive uncertainty. Conversely, questions that remain incorrectly answered by AI even after review are probably the most valuable for evaluating not just the quantity of a student's knowledge but, more importantly, their ability to think critically, reason, and reflect, which should be the cornerstone of any educational assessment. The downside of AI in education is that it can also generate essays or summaries requested by instructors. Although these AI systems can produce high-quality text, their inability to understand content makes them prone to factual errors. Therefore, students must always review and verify any AI-generated material, checking both the accuracy of statements and the validity of references. This issue is clearly reflected in our results: bibliographic support varied widely, from nonexistent in 2 systems to apparently adequate in the other 2, which provided about 3 or 4 references per answer. However, the quality of these references was limited—in Bing, only 7.2% of citations were indexed in PubMed, while in Bard, nearly 90% of its citations were fabricated, and < 1% were PubMed-indexed. Similar findings were reported by Chen *et al.*, who found that 20.6% of AI-generated citations were false,[29] possibly due to the architecture of these systems, which prioritize natural-language fluency over factual accuracy or evidential support.[30]

Response times among the AIs were extremely short, which—along with their insensitivity to human factors such as fatigue or emotional state—is one of their major advantages, facilitating large-scale usability. Although statistically significant time differences were observed among systems, these were clinically irrelevant, as even the slowest responses began within approximately 20 seconds.

Additionally, there were differences in the verbosity of responses: some were very concise, while others were notably verbose. This aspect is of little practical importance, since most AI systems allow users to specify the desired response length, making them adaptable for tasks such as writing summaries or structured essays of predetermined length.

This study has several limitations. Only 30 questions were posed to each AI, all within a single specialty—clinical toxicology—which precludes extrapolation to other areas of emergency or acute care medicine, or to other disciplines. Moreover, these AI systems are evolving rapidly, and it is likely that their performance would improve substantially with newer versions available in the near future.

Conclusions

The AI systems tested demonstrated sufficient capability to successfully pass a clinical toxicology examination, suggesting access to accurate knowledge and the ability to provide coherent, understandable answers. They may therefore be useful tools to complement or enhance clinical care, education, and research in health sciences, including emergency medicine, provided that ethical, moral, and legal considerations are observed. However, their current inability to certify the truthfulness of their output makes expert review indispensable.

Consequently, the implementation of AI systems is likely to extend beyond academic learning into management, clinical, and research applications across all health sciences curricula.

ARTICLE INFORMATION

Conflict of Interest Disclosures: None reported.

Funding: The authors declare the non-existence of funding in relation to this article.

Ethical responsibilities: The authors have confirmed the maintenance of confidentiality and respect for the patient rights, agreement of publication, and transfer of rights to Revista Española de Urgencias y Emergencias.

Article not commissioned by the Editorial Board and with external peer review.

Note of the editors: This is a BOWMAN-generated English translation of the officially indexed Spanish-language article, which should be cited as Rev Esp Urg Emerg. 2024;3:15-19. In this translated version, the editors have supervised the process; however, it cannot be ruled out that some errors resulting from the artificial intelligence translation process may have gone unnoticed.

REFERENCES

1. Chat GPT. Autodefinición de inteligencia artificial. (Accessed 17 Decemeber 2023). Available at:
2. Howard J. Artificial intelligence: Implications for the future of work. Am J Ind Med. 2019; 62:917-26.
3. Pham KT, Nabizadeh A, Selek S. Artificial intelligence and chatbots in psychiatry. Psychiatr Q. 2022;93:249-53.
4. Zawiah M, Al-Ashwal FY, Gharaibeh L, Abu Farha R, Alzoubi KH, Abu Hammour K, et al. ChatGPT and clinical training: perception, concerns, and practice of pharm-d students. J Multidiscip Healthc. 2023;16:4099-110.
5. Tsiknakis N, Trivizakis E, Vassalou EE, Papadakis GZ, Spandidos DA, Tsatsakis A, et al. Interpretable artificial intelligence framework for COVID-19 screening on chest X-rays. Exp Ther Med. 2020;20:727-35.
6. Buelga ML, Ramírez J, Alonso GL. El reloj inteligente *(smartwatch)* ante el bloqueo auriculoventricular completo: un reto por encima de sus posibilidades. Emergencias. 2023;35:478-9.
7. Moreno E, Pueyo I, Sánchez M, Martín M, Masip J. Experiencia de Mediktor®: un nuevo evaluador de síntomas basado en inteligencia artificial para pacientes atendidos en el servicio de urgencias. Emergencias. 2017;29:391-6.
8. Nogué S. Toxicólogo y urgenciólogo: una nueva variante del cangrejo ermitaño. Emergencias. 2009;21:62-4.
9. Brin D, Sorin V, Vaid A, Soroush A, Glicksberg BS, Charney AW, et al. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. Sci Rep. 2023;13:16492.
10. González-Díaz A, Matos-Castro S, Arruabarrena-Urrestarazu N, González-Valladares E, Molina-Padilla S, Ferrer-Dufol A, et al. Evolución de las intoxicaciones agudas por productos químicos en el quinquenio 2015-2019, registradas por el Sistema Español de Toxicovigilancia (SETv). Rev Esp Urg Emerg. 2023;2:30-5.
11. Marcos DN, Cuerpo S, Coll-Vinent B. Coma en paciente *nomen nescio*. Rev Esp Urg Emerg. 2023;2:56-7.
12. Piraianu AI, Fulga A, Musat CL, Ciobotaru OR, Poalelungi DG, Stamate E, et al. Enhancing the evidence with algorithms: how artifi-

cial intelligence is transforming forensic medicine. Diagnostics (Basel). 2023;13:2992.
13. Ayers JW, Zhu Z, Poliak A, Leas EC, Dredze M, Hogarth M, et al. Evaluating artificial intelligence responses to public health questions. JAMA Netw Open. 2023;6:e2317517.
14. Kumar M, Nguyen TPN, Kaur J, Singh TG, Soni D, Singh R, et al. Opportunities and challenges in application of artificial intelligence in pharmacology. Pharmacol Rep. 2023; 75:3-18.
15. Alam M, Hallak JA. AI-automated referral for patients with visual impairment. Lancet Digit Health. 2021;3:e2-e3.
16. Augusto JB, Davies RH, Bhuva AN, Knott KD, Seraphim A, Alfarih M, et al. Diagnosis and risk stratification in hypertrophic cardiomyopathy using machine learning wall thickness measurement: a comparison with human test-retest performance. Lancet Digit Health. 2021;3:e20-e28.
17. Sechopoulos I, Teuwen J, Mann R. Artificial intelligence for breast cancer detection in mammography and digital breast tomosynthesis: State of the art. Semin Cancer Biol. 2021;72:214-25.
18. Soffer S, Morgenthau AS, Shimon O, Barash Y, Konen E, Glicksberg BS, et al. Artificial intelligence for interstitial lung disease analysis on chest computed tomography: A systematic review. Acad Radiol. 2022; Suppl 2:S226-S235.
19. Myszczynska MA, Ojamies PN, Lacoste AMB, Neil D, Saffari A, Mead R, et al. Applications of machine learning to diagnosis and treatment of neurodegenerative diseases. Nat Rev Neurol. 2020;16:440-56.
20. Roman A, Al-Sharif L, Al Gharyani M. The Expanding Role of ChatGPT (Chat-Generative Pre-Trained Transformer) in Neurosurgery: A systematic review of literature and conceptual framework. Cureus. 2023;15:e43502.
21. Fraser K, Bruckner DM, Dordick JS. Advancing predictive hepatotoxicity at the intersection of experimental, in silico, and artificial intelligence technologies. Chem Res Toxicol. 2018; 31:412-30.
22. Rokhshad R, Ducret M, Chaurasia A, Karteva T, Radenkovic M, Roganovic J, et al. Ethical considerations on artificial intelligence in dentistry: A framework and checklist. J Dent. 2023;135:104593.
23. Sonnenschein K, Stojanovic SD, Dickel N, Fiedler J, Bauersachs J, Thum T, et al. Artificial intelligence identifies an urgent need for peripheral vascular intervention by multiplexing standard clinical parameters. Biomedicines. 2021; 9:1456.
24. Nogue-Xarau S, Amigó-Tadin M. Ríos-Guillermo J. ¿Puede la inteligencia artificial ayudar al urgenciólogo en el diagnóstico de las intoxicaciones? Emergencias 2024 (en prensa).
25. Mello MM, Guha N. ChatGPT and physicians' malpractice risk. JAMA Health Forum. 2023;4:e231938.
26. Organización Mundial de la Salud (OMS). Informe sobre Inteligencia Artificial (IA) aplicada a la salud y seis principios rectores relativos a su concepción y utilización. (Accessed 18 Decemeber 2023). Available at: https://www.who.int/es/news/item/28-06-2021-who-issues-first-global-report-on-ai-in-health-and-six-guiding-principles-for-its-design-and-use.
27. Eysenbach G. 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. 2023;9:e46885.
28. 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? The Implications of Large Language Models for Medical Education and Knowledge Assessment. JMIR Med Educ. 2023;9:e45312.
29. Chen A, Chen DO. Accuracy of chatbots in citing journal articles. JAMA Netw Open. 2023;6:e2327647.
30. Ryan DK, Maclean RH, Balston A, Scourfield A, Shah AD, Ross J. Artificial intelligence and machine learning for clinical pharmacology. Br J Clin Pharmacol. 2023 Oct 16. doi: 10.1111/bcp.15930.