REUE | Original article

Toxicology questions in Spanish medical licensing exams (MIR): accuracy of artificial intelligence vs a group of clinical toxicology experts

Santiago Nogué-Xarau[1], José Ríos-Guillermo[2], Montserrat Amigó-Tadin[3], en nombre del Grupo de Trabajo de Toxicología de la Sociedad Catalana de Medicina de Urgencias y Emergencias (SoCMUETox)

OBJECTIVE. To assess the ability of several artificial intelligence (AI) systems to correctly answer toxicology questions from Spain's *Médico Interno Residente* (MIR) licensing exams and to compare their accuracy with that of a group of clinical toxicologists.

MATERIAL AND METHODS. We selected toxicology-related questions from the MIR exams (2019–2023) and showed them to 7 AI chatbots (ChatGPT, Gemini, Copilot, Luzia, Claude, Deepseek, and Le Chat) and to a group of clinical toxicologists. The number of correct answers was recorded for each participant.

RESULTS. A total of 44 questions were included. AI systems completed the exam in a median of 1.01 (0.82–1.52) minutes vs 42.00 (28.50–53.50) minutes for toxicologists ($P < .001$). AI achieved a median of 41 (39–42) correct answers while toxicologists achieved 32 answers (26–36) ($P < .001$). No differences were found among toxicologists by age, sex, or specialty, nor between theoretical and case report-based questions.

CONCLUSIONS. AI chatbots answered toxicology questions from MIR exams faster and with higher accuracy than a group of clinical toxicologists.

Keywords: Clinical toxicology. Resident physician. MIR exam. Toxicology exam. Artificial intelligence.

Preguntas de toxicología en los exámenes MIR: aciertos en las respuestas de las inteligencias artificiales en comparación con un grupo de expertos en toxicología clínica

OBJETIVO. Valorar la capacidad de varios sistemas de inteligencia artificial (IA) para acertar la respuesta correcta a preguntas de toxicología que se han formulado en las convocatorias del Ministerio de Sanidad para obtener plaza de médico-interno-residente (MIR), y compararla con los aciertos que a las mismas preguntas ha ofrecido un grupo de expertos en toxicología clínica.

MATERIAL Y MÉTODOS. Se revisaron los cuestionarios de los exámenes MIR de las cinco últimas convocatorias (2019-2023) y se seleccionaron las preguntas de toxicología. Estas preguntas se pasaron a siete *chatbots* de IA (ChatGPT, Gemini, Copilot, Luzia, Claude, Deepseek y Le Chat), solicitando las respuestas correctas. Las mismas preguntas se formularon también a un grupo de toxicólogos.

RESULTADOS. Se incluyeron 44 preguntas. El tiempo cronometrado para completar este examen por las IA fue de 1,01 (0,82-1,52) minutos, mientras que el tiempo estimado que precisaron los toxicólogos fue de 42 (28,5-53,5) minutos ($p < 0,001$). Las IA acertaron una mediana de 41 (39-42) respuestas, mientras que la mediana de respuestas acertadas por el grupo de toxicólogos fue de 32 (26-36) ($p < 0,001$). En el grupo de toxicólogos no se encontraron diferencias en los aciertos en función de su edad, sexo o especialidad médica. Tampoco se han encontrado diferencias en los aciertos entre preguntas teóricas y casos clínicos.

CONCLUSIONES. Los *chatbots* de IA contestan las preguntas de toxicología de los exámenes MIR más rápidamente y obtienen mejores resultados que un grupo de toxicólogos.

Palabras clave: Toxicología clínica. Médico Residente. Examen MIR. Examen de toxicología. Inteligencia artificial.

Introduction

In Spain, as of today, there are 48 medical specialties officially recognized by the Ministry of Health,[1] and the most recent among them is Emergency and Urgent Care Medicine (EUCM).[2,3] Although Clinical Toxicology (CT) is not an officially recognized medical specialty in Spain, it is

Author Affiliations: [1]Fundación Española de Toxicología Clínica, Spain. [2]Departamento de Farmacología Clínica, Hospital Clínic y Unidad de Estadística Médica, Instituto de Investigaciones Biomédicas August Pi i Sunyer (FCRB-IDIBAPS), Barcelona, Spain. [3]Servicio de Urgencias, Hospital Clínic de Barcelona, Spain.
Corresponding Author: Santiago Nogué-Xarau. Fundación Española de Toxicología Clínica. Barcelona, Spain.
E-mail: snoguex@gmail.com
Article Information: Received: 4-8-2025. Accepted: 8-10-2025. *Online:* 22-10-2025.
Editor in Charge: Guillermo Burillo-Putze.

acknowledged as a specialty or subspecialty in other countries (for example, the United States). In the absence of such recognition, and given the large number of poisonings that occur and the extensive care provided for these conditions in emegerncy departments (EDs)[4] and intensive care units (ICUs),[5] CT has been incorporated into the training curricula of several specialties, such as Internal Medicine, Intensive Care Medicine, Occupational Medicine, and, naturally, EUCM.

The process of becoming a medical specialist within the Spanish healthcare system involves a competitive national examination organized annually by the Spanish Ministry of Health, which grants access to a residency training position. Medical graduates are eligible to participate in this process. In this examination, candidates for residency positions (MIR) complete a multiple-choice test of approximately 200 questions. Together with the evaluation of their academic record, this yields a final score and ranking number, which allows candidates to choose both their specialty and the hospital where they will train for 4–5 years before obtaining specialist certification.

These examinations assess both medical knowledge and the applicant's clinical reasoning skills. In recent years, and particularly since November 2022, when ChatGPT, an advanced artificial intelligence (AI) chatbot model designed to understand and generate text, became widely available, considerable speculation has arisen regarding the usefulness of AI in medicine. AI systems have been subjected to numerous evaluations, including their ability to assist in the diagnosis and treatment of vascular surgery patients,[6] their role in radiologic diagnosis,[7] their potential to interpret electrocardiograms, and their capacity to predict patient outcomes in EDs.[8]

The primary endpoint of this study was to test the ability of several AI systems to correctly answer questions related to a subspecialty of EUCM—namely Clinical Toxicology—that appeared in the MIR examinations over the past five years. As a secondary endpoint, we sought to compare the responses generated by these AI systems with those provided by a group of toxicology experts (GET).

Materials and methods

We obtained the MIR examinations pertaining to the 5-year period 2019–2023 from the official website of the Spanish Ministry of Health.[9] Each examination consisted of multiple-choice questions, with four possible answers per question, of which only one was correct. The same website provided the official correct answers.

Two of the authors (SNX and MAT), both with extensive experience in CT, reviewed all the questions from these examinations (185 in 2019 and 210 per year from 2020 onward) and selected those related to toxicology (clinical cases and knowledge-based questions concerning acute or chronic poisoning). Questions related to adverse drug reactions were also included, while all image-based questions were excluded. The selected questions were compiled into a Word document with their corresponding answer options. Each question was classified as either case-based or theoretical.

To assess CT knowledge, seven widely used AI chatbot systems were selected: ChatGPT (OpenAI, version 4; https://chatgpt.com), Gemini (Google, version 1.5-Flash; https://gemini.google.com/app?hl=es), Copilot(Microsoft, version not specified; https://copilot.microsoft.com), Luzia (version not specified; https://www.luzia.com), Claude (Anthropic, version 3.5 Haiku; https://claude.ai), DeepSeek (version V3; http://deepseek.com), and Le Chat (Mistral AI, version not specified; https://chat.mistral.ai). All were free-access versions available via their respective websites. Throughout January 2025, the following prompt was entered into each AI system: "For the questions I am about to provide, I need you to tell me which of the 4 answer options is correct for each question."The 44 selected questions were then entered verbatim, in groups of 9, 9, 9, 9, and 8 questions. The time required by each AI system to complete all responses was recorded.

Simultaneously, collaboration was requested from the Toxicology Working Group of the Catalan Society of Emergency and Urgent Care Medicine (SoCMUETox), one of the societies within the *Acadèmia de Ciències Mèdiques de Catalunya i de Balears*. This group has a well-established clinical, educational, and research background in both CT and EUCM. Members who agreed to participate received the same document by email and were asked to indicate the correct answers without consulting any bibliographic or external sources. Thirty GET members participated: 27 physicians and 3 nurses, all working in hospital-based or prehospital EDs or ICUs. The time required by these participants to complete the questionnaire was calculated using the "editing time" metadata available in the Word document.

For both groups, the number of correct, incorrect, and blank responses was recorded. Following the official MIR scoring system, each correct answer was awarded three points, each incorrect answer deducted one point, and blank answers received zero points. The seven AI systems were compared with each other, and their results were also compared with those of the GET.

Statistical analysis was performed using SPSS (version 26, Armonk, IBM Corp., New York, NY, United States). In addition to comparing correct, incorrect, and blank responses, the rate of correct answers was calculated, considering incorrect and blank answers as incorrect. This allowed direct comparison between the AI systems and the GET, as chatbots are programmed to always provide an answer and never leave questions blank. The number of correct responses was expressed as median and 25^{th}–75^{th} percentiles. Percentages of correct answers were also calculated for the AI systems.

Within the GET, comparisons were made according to participant gender, age (born before or after 1977), profession (physician or nurse), and whether the specialty had been obtained through the MIR system. Differences were analyzed using the Mann–Whitney U test, and for specialty categories (family medicine, internal medicine, or other)

the Kruskal–Wallis test was used. When statistically significant differences were found, post-hoc pairwise comparisons were performed using the Mann–Whitney U test. Comparisons between AI systems and expert evaluators were also conducted using the Mann–Whitney U test. A *P*-value < .05 was considered statistically significant.

Results

A total of 44 questions from the MIR examinations were included: 9 from the 2019 examination, 6 from 2020, 8 from 2021, 10 from 2022, and 11 from 2023. Of these 44 questions, 23 (52.3 %) belonged strictly to the field of Clinical Toxicology, while 21 (47.7 %) were related to adverse drug reactions.

Table 1 shows the time required to complete the toxicology responses, which was significantly shorter for the AI systems compared with the expert group (*P* < .001). The accuracy rate and total scores obtained by the AI systems and the experts were also compared, revealing superior performance by the AI systems (*P* < .001).

Table 2 presents the accuracy of responses from members of the SoCMUETox group according to participant gender, age, profession, specialty, and whether the specialty had been obtained through the MIR pathway. No statistically significant differences were observed for any of these variables.

Differences in accuracy among the various AI systems and between the AI systems and the expert group were also analyzed (Table 3). Copilot and Le Chat achieved the highest scores (95.5 % correct answers), although no statistically significant differences were found among the seven AI systems evaluated. The median number of correct answers in the expert group was 32 (72.7 %), which was significantly lower than that of the AI systems (*P* < .001).

Additionally, the impact of question type was evaluated (Table 4): 24 case-based questions and 20 theoretical questions were analyzed. No significant differences in accuracy were found between the two question types, either among the AI systems or within the expert group.

Table 1. Responses provided by artificial intelligence systems to MIR examination questions compared with the SoCMUETox Expert Group*

Variable	SoCMUETox Group* (n = 30) Median [25th–75th percentile]	AI** (n = 7) Median [25th–75th percentile]	*P* value
Toxicology questions (n = 44)			
Time to complete the 44 toxicology responses (minutes)	42 [28.5-53.5]	1.01 [0.82-1.52]	< .001
Number of correct answers	32 [26-36]	41 [39-42]	< .001
Number of incorrect answers	7.5 [4-10]	3 [2-5]	
Number of unanswered questions	3.5 [0-6]	0 [0]	
Total score***	92 [68-101]	120 [112-124]	< .001

*SoCMUETox: Toxicology Working Group of the Catalan Society of Emergency and Urgent Care Medicine.
**AI: Artificial Intelligence.
***Each correct answer scores +3 points, each incorrect answer scores –1 point, and unanswered questions receive 0 points.

Discussion

AI tools based on generative language models, such as ChatGPT, have demonstrated remarkable knowledge in health sciences when provided with well-formulated prompts. This capability has allowed them, for example, to successfully pass medical licensing examinations in the United States[10] and to perform well in selection examinations for emergency physicians across certain Spanish regions.[11]

In the EUCM field, recent reviews confirm the potential of large language models (LLMs) to influence emergency medicine by improving clinical decision-making, optimizing workflows, and enhancing patient outcomes.[12,13] However, these systems have not yet been widely implemented in routine clinical practice. Effective integration of LLMs will require collaborative efforts and rigorous evaluation to ensure safe and effective clinical application.[14,15] In pediatric emergency medicine, experiences also suggest that LLMs could be used to enhance specialist training.[16]

Within CT, ChatGPT has demonstrated its ability to diagnose real cases of organophosphate poisoning,[17] provide useful clinical information in venomous snakebite cases,[18] and generate coherent summaries of real calls to national poison information centers, thereby improving operational efficiency.[19]

To our knowledge, this study is the first in Spain to compare 7 AI systems using MIR examination questions limited to the domain of CT. Their accuracy was remarkably high and significantly superior to that of a group of experts who manage poisoned patients daily. Previous work from our group showed that these AI systems also perform very well in toxicology examinations at the School of Medicine of *Universidad de Barcelona* (Barcelona, Spain)[20] and can generate toxicology responses that are indistinguishable from those of human experts according to the Turing test.[21]

Table 2. Differences by sex, age, profession, and specialty of the SoCMUETox Group* in accuracy on 44 toxicology questions from the MIR examination

Evaluator	Correct toxicology answers [Median (SD)]	*P* value
Sex		.278
Female (n = 19)	32 (5.16)	
Male (n = 11)	29.64 (6.39)	
Age		.652
Born < 1977 (n = 16)	30.69 (5.83)	
Born ≥ 1977 (n = 14)	31.64 (5.62)	
Profession		.153
Physician (n = 27)	31.63 (5.48)	
Nurse (n = 3)	26.67 (6.35)	
Specialty		.363
Family and Community Medicine (n = 10)	30 (6.07)	
Internal Medicine (n = 5)	33.8 (2.86)	
Other specialties or none (n = 15)	33.3 (7.26)	
Specialty obtained via MIR training		.507
Yes (n = 19)	31.16 (5.8)	
No (n = 11)	29.2 (5.72)	

*SoCMUETox: Toxicology Working Group of the Catalan Society of Emergency and Urgent Care Medicine.
MIR: Medical Intern–Resident.

Table 3. Correct and incorrect answers provided by AI systems and by the clinical toxicology expert Group to Toxicology questions from the 2019–2023 MIR examinations

Evaluator	Response time* (min)	*P* value	Correct answers n (%)	Incorrect answers n (%)	Unanswered n (%)	Total Score	*P* value
Toxicology questions (n = 44)							
Copilot	0.82	< .001**	42 (95.5)	2 (4.5)	0	124	.278**
ChatGPT	2.63		41 (93.2)	3 (6.8)	0	120	
Luzia	0.88		39 (88.6)	5 (11.4)	0	112	
Gemini	0.77		37 (84.1)	7 (15.9)	0	104	
Claude	1.01		40 (90.9)	4 (9.1)	0	116	
DeepSeek	1.52		41 (93.2)	3 (6.8)	0	120	
Le Chat	1.14		42 (95.5)	2 (4.5)	0	124	
Group of 30 experts [median (IQR)]	42 (28-54)	< .001***	32 (25-36)	7.5 (4-11)	4 (0-8)	88 (65-100)	< .001***

*For AI systems, response time refers to the time elapsed from prompt submission until completion of all 44 answers. For human evaluators, response time refers to the editing time of the Word document while it remained open.
**Comparison across 7 AI systems.
***Comparison between the seven AI systems and the group of 30 experts.
AI: Artificial Intelligence.

Although these systems were not specifically trained in toxicology, their access to extensive knowledge and near-instantaneous response generation makes them potentially useful as support tools for clinicians managing poisoned patients in EDs. Nevertheless, the present study design does not allow us to conclude that AI systems improve real-world patient care—this will require dedicated clinical trials. Furthermore, these systems may generate erroneous responses, as they themselves acknowledge, meaning that diagnostic and therapeutic suggestions must always be critically evaluated by the clinician responsible for patient care.[22]

While no statistically significant differences were observed among the seven AI models, slight variability was noted, with Copilot and Le Chat leading in performance. None achieved 100 % accuracy. This underscores the indispensable role of human decision-making: the physician remains the ultimate authority in clinical decisions. These chatbots complement but do not replace traditional evidence-based medical literature. Notably, the AI systems used in this study were open-access models with unrestricted internet access. Closed AI systems, trained exclusively on curated clinical content without internet connectivity, may perform differently and warrant evaluation in future studies.

The expert group achieved lower scores than the AI systems. This difference may reflect the fact that experts answered spontaneously without consulting reference material, whereas LLMs leverage massive information repositories. Additionally, expert responses may have been influenced by fatigue or variable working conditions—factors that do not affect AI systems. Despite their extensive clinical experience, human expertise could not fully compensate for the immediate, expansive knowledge accessible to LLMs. No differences in accuracy were observed among experts according to sex, age, or profession, suggesting a homogeneous knowledge base within the group.

Although no statistically significant differences were found in the performance of the AI systems among the seven models evaluated, slight variability was observed, with Copilot and Le Chat leading in the number of correct answers for this type of question, although none of them reached 100 % accuracy. This finding underscores the importance of the final human decision in patient care, with the physician always remaining the ultimate authority responsible for clinical decisions. These chatbots complement but do not replace traditional literature searches for evidence-based medical information. In this study, open-access AI systems with unrestricted internet access and the use of diverse information sources were employed. By contrast, closed AI systems, trained exclusively on a restricted clinical corpus compiled from manuals and official guidelines, also exist. Future studies should evaluate these closed AI models, which lack internet access and rely solely on curated clinical content to determine whether AI functions as a reliable clinical tool or merely as an advanced search engine.

The Expert Group (GET) achieved lower results than the AI systems. This difference may be explained by several factors. First, group members were asked to answer spontaneously and without consulting any bibliographic or medical information sources, in contrast to the vast volume of information available to LLMs, which are explicitly designed to access and integrate such data. Furthermore, the questionnaires may have been completed by human participants under highly heterogeneous conditions, such as at the end of a workday or during clinical duties, circumstances associated with physical or mental fatigue that may reduce performance—fatigue that does not affect LLMs. Although toxicologists contribute clinical knowledge grounded in science and personal experience, this was insufficient to counterbalance the immediate and nearly unlimited access to online information available to LLMs. Within the expert group, no differences were observed in the rate of correct answers to toxicology questions according to sex, age, or profession, suggesting a relatively homogeneous level of expertise within the group.

If the toxicology results obtained by the AI systems were extrapolated to the remainder of the MIR examination, all AI models would have passed with exceptionally high scores. For example, in the 2021 MIR examination,

Table 4. Answers provided by AI systems vs those of the SoCMUETox expert group, according to whether the toxicology question was theoretical or case-based*

Scientific area	Question type	Outcome	SoCMUETox Group (n = 30) Median [25th–75th percentile]	AI systems (n = 7) Median [25th–75th percentile]	*P* value
Toxicology questions (n = 44)					
	Theoretical question (n = 24)	Correct answers	21 [19-22]	22 [21-22.5]	.529
		Incorrect answers***	3 [2-5]	2 [2-2.5]	
	Clinical case (n = 20)	Correct answers	17 [15-18]	18 [17-18]	.337
		Incorrect answers***	3 [2-5]	2 [2-3]	

*SoCMUETox: Toxicology Working Group of the Catalan Society of Emergency and Urgent Care Medicine.
**AI: Artificial Intelligence.
***For human evaluators, unanswered questions were considered incorrect. AI systems provided no unanswered responses.

considering only the number of correct, incorrect, and blank responses, the AI systems would have ranked as follows: 1st (Copilot and Le Chat), 3rd (ChatGPT and DeepSeek), 6th (Claude), 10th (Luzia), and 75th (Gemini) among 11,829 candidates.[23] Applying the same extrapolation to the 2019 examination, the rankings would have been: 4th (Copilot and Le Chat), 167th (ChatGPT and DeepSeek), 359th (Claude), 4,109th (Luzia), and 7,588th (Gemini) among 11,827 candidates.[24] No reasonable explanation was found for the poorer AI performance relative to human candidates in 2019, except for the possibility that examinees were better prepared that year compared with 2021.

Several limitations of this study should be acknowledged. First, question selection was based on the judgment of two experts, potentially introducing selection bias. In addition, because the number of pure clinical toxicology questions identified in the MIR examinations was small (n = 23), questions related to adverse drug reactions (n = 21) were included, as the boundary between toxic and idiosyncratic drug reactions is often blurred. Consequently, the evaluated knowledge cannot be considered exclusive to toxicology in the strictest sense. Furthermore, the small total number of questions included in the study (4.4 % of the 1,000 questions analyzed) limits the generalizability of the findings. The extrapolation used to estimate AI ranking in the MIR examination is therefore illustrative rather than definitive. All AI systems were evaluated using free-access versions, and it remains unknown whether subscription-based models would have yielded even better performance. The number of human participants was also limited and drawn predominantly from a single Spanish region (Catalonia), which restricts representativeness. Finally, while AI response times were precisely measured, human response times were estimated from the duration of Word document editing, which does not guarantee continuous engagement with the task.

Conclusions

AI systems achieved excellent scores on toxicology questions from the MIR examinations, outperforming a group of clinical toxicology experts and generating their responses significantly faster.

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.

Data Availability: Data are available upon request from the corresponding author.

Author Contributions (CRediT): SNX: conceptualization, investigation, methodology, writing (original draft). JRG: formal analysis, methodology, writing (review and editing). MAT: conceptualization, investigation, writing (review and editing).

Use of Generative Artificial Intelligence Tools: The authors declare that no generative AI tools were used in the preparation of this article.

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. 2026;5:36-41. 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.

ADENDUM

Toxicology Working Group of the Catalan Society of Emergency and Urgent Care Medicine (SoCMUETox) who participated in this study: África de la Cruz Ramos, Alberto Moreno Destruels, Alma María Palomino, Àngels Gispert Ametller, August Supervía Caparrós, Carlos García Gutiérrez, Conxita Moll Tudurí, Cristina Ramió Lluch, Elena Fuentes González, Emilio Salgado García, Evangelos Papoutsidakis, Francisca Córdoba Ruiz, Héctor Hernández Ontiveros, Indalecio Morán Chorro, Irina Hernández Medina, Jordi Puiguriguer Ferrando, Josep Piqueras Carrasco, Laia Ferrer Caballé, Laia Guerrero González, Lidia García Gibert, Lidia Martínez Sánchez, Lluisa Corral Ansa, María Codinach Martín, Marinela Guzmán Carvajal, Marta Serrano Giménez, Miguel Galicia Paredes, Rosana Muñoz Bermúdez, Vicenç Ferrés Padró y Yolanda Ibañez Borau.

REFERENCES

1. Ministerio de Sanidad. Especialidades en Ciencias de la Salud. (Accessed 25 November 2024). Available at: https://www.boe.es/eli/es/rd/2008/02/08/183.
2. Ministerio de Sanidad. Especialidad de Medicina de Urgencias y Emergencias. (Accessed 25 November 2024). Available at: https://www.sanidad.gob.es/areas/profesionesSanitarias/profesiones/tituloEspecialista/home.htm.
3. Vázquez Lima MJ. La Especialidad de Medicina de Urgencias y Emergencias ya es una realidad en Spain. Emergencias. 2024;36:321-3.
4. Supervía A, Córdoba F, Ruíz Antorán B, Martín Pérez B, Martínez Baladrón A, Urdangarín A et al. Registro EPITOX de intoxicaciones en Spain. Año 2024. Rev Esp Urg Emerg. 2025;4:210-7.
5. Socías A, Nogué S, Alcaraz RM, Morán I, Montero FJ, Palomar M, et al. Evolución de las intoxicaciones en las unidades de cuidados intensivos españoles. Comparación de 2 períodos. Med Intensiva. 2021;45:e4-e6.
6. Alexiou VG, Sumpio BE, Vassiliou A, Kakkos SK, Geroulakos G. Artificial Intelligence in Diagnosing and Managing Vascular Surgery Patients: An Experimental Study Using the GPT-4 Model. Ann Vasc Surg. 2024;23:S0890-5096(24)00705-2.
7. Strubchevska O, Kozyk M, Kozyk A, Strubchevska K. The Role of Artificial Intelligence in Diagnostic Radiology. Cureus. 2024;16:e72173.
8. Zaboli A, Brigo F, Ziller M, Massar M, Parodi

M, Magnarelli G, et al. Exploring ChatGPT's potential in ECG interpretation and outcome prediction in emergency department. Am J Emerg Med. 2024;88:7-11.
9. Ministerio de Sanidad. Cuadernos de exámenes para optar a la formación médica especializada. (Accessed 3 December 2024). Available at: https://fse.mscbs.gob.es/fseweb/view/public/datosanteriores/cuadernosExamen/busquedaConvocatoria.xhtml.
10. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment. JMIR Med Educ. 2023;9:e45312.
11. García-Rudolph A, Sanchez-Pinsach D, Opisso E. ChatGPT's performance in the Specialist Health Practitioner exam for Hospital Emergency, responses from GPT-3.5 and GPT-4.0 to 150 multiple-choice questions. Eur J Emerg Med. 2024;31:438-9.
12. Preiksaitis C, Ashenburg N, Bunney G, Chu A, Kabeer R, Riley F, et al. The Role of Large Language Models in Transforming Emergency Medicine: Scoping Review. JMIR Med Inform. 2024;12:e53787.
13. Sáenz-Abad D, Sachi Martínez-Mihara M, Lahoza-Pérez MC. La inteligencia artificial como herramienta de apoyo diagnóstico en urgencias. Emergencias. 2025;37:78-9.
14. Romero Olóriz C. Inteligencia artificial en incidentes con múltiples víctimas: estado actual y perspectivas. Emergencias. 2025;37:159-60.
15. González-Martínez F, Garrido NJ, Mateo J. Inteligencia artificial en la práctica clínica de urgencias: más realidad que fascinación. Emergencias. 2025;37:159-60.
16. Ramgopal S, Varma S, Gorski JK, Kester KM, Shieh A, Suresh S. Evaluation of a Large Language Model on the American Academy of Pediatrics' PREP Emergency Medicine Question Bank. Pediatr Emerg Care. 2024;40:871-5.
17. Sabry Abdel-Messih M, Kamel Boulos MN. ChatGPT in Clinical Toxicology. JMIR Med Educ. 2023;9:e46876.
18. 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.
19. Matsler N, Pepin L, Banerji S, Hoyte C, Heard K. Use of large language models to optimize poison center charting. Clin Toxicol (Phila). 2024;62:385-90.
20. 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.
21. Nogué-Xarau S, Ríos-Guillermo J, Amigó-Tadin M, Grupo SoCMUETox. Comparación de las respuestas a preguntas sobre intoxicaciones generadas por sistemas de inteligencia artificial y las creadas por toxicólogos clínicos. Emergencias. 2024;36:351-8.
22. Carballo Cardona C, Iglesias Sigüenza A, Deza Palacios R, Soriano Arroyo R, Rodríguez Fuertes P, Tejada Sorados RM, et al. Inteligencia artificial en medicina de urgencias y emergencias: ¿amenaza tecnológica o aliado clínico?. Rev Esp Urg Emerg. 2026;5:59-64.
23. Relación provisional de los resultados de las pruebas selectivas, año 2021, Medicina. (Accessed 27 February 2025). Available at: https://www.consalud.es/uploads/s1/18/38/94/2/resultados-provisionales-mir-2022.pdf.
24. Relación provisional de los resultados de las pruebas selectivas, año 2019, Medicina. (Accessed 12 February 2025). Available at: https://blog.promir.es/wp-content/uploads/2020/02/Listado-provisional-de-resultados-de-las-pruebas-selectivas-del-MIR-2019.pdf.