Article

Not peer-reviewed version

Evaluating ChatGPT Efficacy in Navigating the Spanish Medical Residency Entrance Examination (MIR): A New Horizon for AI in Clinical Medicine

Francisco Guillen-Grima *, Sara Guillen-Aguinaga , Laura Guillen-Aguinaga , Rosa Alas-Brun , Luc Onambele , Wilfrido Ortega , Rocio Montejo , Enrique Aguinaga-Ontoso , Paul Barach , Ines Aguinaga-Ontoso *

Posted Date: 19 September 2023

doi: 10.20944/preprints202309.1272.v1

Keywords: Machine learning; Artificial inteligence; ChatGPT; GPT 3.5; GPT-4; Medical education; quality of care

Article

Evaluating ChatGPT Efficacy in Navigating the Spanish Medical Residency Entrance Examination (MIR): A New Horizon for AI in Clinical Medicine

Francisco Guillen-Grima [1,2,3,4,*], Sara Guillen-Aguinaga [1], Laura Guillen-Aguinaga [1,5], Rosa Alas-Brun [1], Luc Onambele [6], Wilfrido Ortega [7], Rocio Montejo [8,9], Enrique Aguinaga-Ontoso [10], Paul Barach [11,12,13] and Ines Aguinaga-Ontoso [1,2,*]

[1] Dept. of Health Sciences, Public University of Navarra; Pamplona, 31008 Spain; ines.aguinaga@unavarra.es (ORCID 0000-0002-2882-930X); saraguillen.sg@gmail.com (ORCID 0000-0003-4748-9520); rosamaria.alas@unavarra.es (ORCID 0000-0003-3450-9342); f.guillen.grima@unavarra.es (ORCID 0000-0001-9749-8076)

[2] Healthcare Research Institute of Navarra (IdiSNA) 31008 Pamplona, Spain.

[3] Department of Preventive Medicine, Clínica Universidad de Navarra, 31008 Pamplona, Spain

[4] CIBER in Epidemiology and Public Health (CIBERESP), Institute of Health Carlos III, 46980 Madrid, Spain

[5] Department of Nursing, Suldal sykehjem, Sands, Norway; guillen.124514@e.unavarra.es (ORCID 0000-0001-7594-6755)

[6] School of Health Sciences, Catholic University of Central Africa, Yaoundé, Cameroon; onambele.luc@ess-ucac.org, (ORCID 0000-0003-1792-4990)

[7] Department of Surgery, Medical and Social Sciences, University of Alcala de Henares, 28871 Alcalá de Henares, Spain. wilfrido.ortega@edu.uah.es (ORCID 0000-0001-5150-8937)

[8] Department of Obstetrics and Gynecology, Institute of Clinical Sciences, University of Gothenburg.413 46 Gothenburg, Sweden; rocio.montejo.rodriguez@gu.se (ORCID 0000-0001-8917-1882)

[9] Department of Obstetrics and Gynecology, Sahlgrenska University Hospital, 413 46 Gothenburg, Sweden.

[10] Department of Sociosanitary Sciences, University of Murcia, Murcia, Spain aguinaga@um.es (ORCID 0000-0002-7994-3559)

[11] Jefferson College of Population Health, Thomas Jefferson School of Medicine, Philadelphia, PA 19107, USA. Paul.Barach@jefferson.edu (ORCID 0000-0002-7906-698X)

[12] Interdisciplinary Research Institute for Health Law and Science, Sigmund Freud University, 1020 Vienna, Austria

[13] Department of Surgery, Imperial College SW7 2AZ, London, UK

[*] Correspondence: f.guillen.grima@unavarra.es (F.G.-G.) Facultad de Ciencias de la Salud UPNA, Avda. de Barañáin sn 31008, Pamplona, Spain; ines.aguinaga@unavarra.es (I.A.-O.) Facultad de Ciencias de la Salud UPNA, Avda. de Barañáin s/n 31008, Pamplona, Spain

Abstract: The rapid progress in artificial intelligence, machine learning, and natural language processing has led to the emergence of increasingly sophisticated large language models (LLMs) enabling their use in healthcare. The study assesses the performance of two LLMs: the GPT-3.5 and GPT-4 models in passing the medical examination for access to medical specialist training in Spain MIR. Our objectives included gauging the model's overall performance, analyzing discrepancies across different medical specialties, discerning between theoretical and practical questions, estimating error proportions, and assessing the hypothetical severity of errors committed by a physician. We studied the 2022 Spanish MIR examination after excluding those questions requiring image evaluations or having acknowledged errors. The remaining 182 questions were presented to the LLM ChatGPT4 and GPT-3.5 in Spanish and English. Logistic regression models analyzed the relationships between question length and question sequence d performance. GPT-4 outperformed GPT -3.5, scoring 86.81% in Spanish ($p<0.001$). English translations had a slightly enhanced performance. Among medical specialties, GPT-4 achieved a 100% correct response rate in several areas, with specialties like Pharmacology, ICU, and Infectious Diseases showing lower performance. The error analysis revealed that while a 13.2% error rate existed, gravest categories like "error requiring intervention to sustain life" and "error resulting in death" had a 0% rate. Conclusions: GPT-4 performs robustly on the Spanish MIR examination, varying its capability to discriminate knoweldge across specialties. While the model's high success rate is commendable, understanding the error severity is critical, especially when considering AI's potential role in real-world medical practice and its implication on patient safety.

Keywords: machine learning; artificial inteligence; ChatGPT; GPT 3.5; GPT-4; medical education; quality of care

1. Introduction

The advances in artificial intelligence, especially in natural language processing, have ushered in new opportunities. Since its appearance on Nov. 30, 2021, ChatGPT has resulted (as of Jul. 23, 2023) in more than 1,096 scientific articles indexed on Scopus, of which 26% are related to health sciences [1]. ChatGPT refers to the conversational interface built upon OpenAI's Generative Pre-trained Transformer (GPT) large language models (LLM), designed to engage in natural language interactions such as GPT-3.5 and GPT-4. While the term "ChatGPT" is used generically in this paper to denote the conversational capabilities of these architectures, we will specify either GPT-3.5 or GPT-4 when discussing attributes or findings related to a particular LLM version. The free version uses the GPT-3.5 model. The premium version that employs GPT-4 is recommended [2]. GPT-4 was developed by self-training to forecast subsequent sentence words by intermittently concealing input words. ChatGPT models have shown a diverse range of applicability, but their performance in specialized medical examinations remains an area of deep interest. While predicting upcoming words is relevant for language creation, it is not directly applicable to diverse health datasets like physiological waveforms due to the complexity and depth of understanding needed to decipher and infer actions in medical decision-making [3].

ChatGPT has shown promise in various medical fields, including allergology, dermatology, and radiology [2,4,5], and in a pool of questions formulated by physicians of 17 specialties [6]. Nevertheless, the performance has not been consistently good. GPT-4 failed to pass the American College of Gastroenterology Self-Assessment Test, failing to pass the examination [7].

ChatGPT has been tested with real questions that patients ask physicians [8]. It has been tested in the Medical Licensing Examination of the United States [9,10], the German State Examination in Medicine [11], the China National Medical Licensing Examination, and the China National Entrance Examination for Postgraduate Clinical Medicine Comprehensive Ability [12], Taiwan's Examination for Medical Doctors [13], and the Japanese Medical Licensing Examination [14]. A meta-analysis of examinations from several medical specialties and several countries found an overall performance of ChatGPT of 61.1% (95% CI 56.1%–66.0%), but this was made with chat GPT 3.5. [15]

The primary aim of this study is to critically assess the proficiency of the LLM GPT-3.5 and GPT-4 models in passing the MIR medical examination. The efficacy of LLM in navigating specialized medical examinations, such as the MIR ("Médico Interno Residente") medical examination in Spain, is unknown. The MIR exam serves as the gateway to medical specialist training in Spain. This rigorous test evaluates candidates' knowledge through a multiple-choice questionnaire, which primarily aims to determine a priority competency score for selecting a specialty and hospital.

We endeavor to gauge the overall performance, delve into potential performance variations across distinct medical specialties, and distinguish the LLM capabilities in handling theoretical versus practical questions. An integral part of this investigation also involves estimating the proportion of errors in the LLM responses. Additionally, we are keen to discern the potential repercussions of such errors, imagining a scenario where a practicing physician might have committed them.

Several hypotheses underpin this research. Firstly, we anticipate that the models might exhibit enhanced aptitude in resolving theoretical questions compared to practical ones. Furthermore, we predict that the more recent GPT-4 model will surpass GPT-3.5 in performance. Differences in performance across various medical specialties are also expected. Intriguingly, we hypothesize that the sequence in which the questions are presented might influence the models' outputs, possibly attributed to model "fatigue." Our research also considers linguistic nuances, postulating that the models' performance may be more optimized for English questions than ones in Spanish. Lastly, the length of the questions might emerge as a significant factor influencing the quality and accuracy of the LLM's responses.

2. Materials and Methods

2.1. Context—The Graduate Medication Education System in Spain

In Spain, there are 47 acknowledged medical specialties. Obtaining such a specialty necessitates spending 4 to 5 years as a Medical Intern (MIR) at a hospital or an accredited health center [16] In 2023, 8,550 specialty positions were made available across Spain, most of which - 8,316 - were situated in public hospitals and health centers, while the remaining 234 positions were offered in private health systems..

The MIR exam was taken in 2023 by 11,578 doctors, comprised of 8,685 Spanish and 2,893 foreign practitioners. Interestingly, foreigners were granted 16.37% of the MIR positions in 2023. Specifically, 1,378 doctors from non-European Union nations secured a job through the MIR exam during the same year [17].

2.2. The Spanish Medical Intern Examination- MIR exam

The MIR exam comprises 200 multiple-choice questions with four potential answers—only one of which is correct. In addition, there are ten reserve questions to address issues related to question formulation or typographical errors, bringing the total to 210 questions. The exam does not adhere to a specified official syllabus; it may include questions on any aspect of medicine and typically draws from topics found in commonly used medical textbooks.

The exam duration is 4 hours and 30 minutes, and it is administered in venues rented by the Ministry of Health throughout Spain. The score from the exam contributes to 90% of the final grade for specialty placement, with the remaining 10% coming from the candidate's academic record. Based on these grades, candidates are ranked, subsequently determining the selection of specialty training spots

We obtained the questions of the 2022 Spanish Medical Residency Entrance Exam Examination (MIR) from the Spanish Ministry of Health web. https://fse.mscbs.gob.es/fseweb/view/public/datosanteriores/cuadernosExamen/busquedaConvocatoria.xhtml . The MIR examination has a master version (Type 0) and four examination types (1 to 4). All the versions of the examinations have the same questions, but the order of questions and answers may differ. From the five available examinations on the website of the Ministry of Health, we chose type 0. We eliminated all questions that required an evaluation of images, like an MRI or an ECG. We also eliminated those questions challenged by the doctors who sate for the exam and were accepted by the Ministry of Health, generally because of errors in the wording or because there were several positive answers. We also included the reserve questions. This left us with 182 questions. Only those questions that required reading of a text were included. The flow diagram is presented in Figure 1.

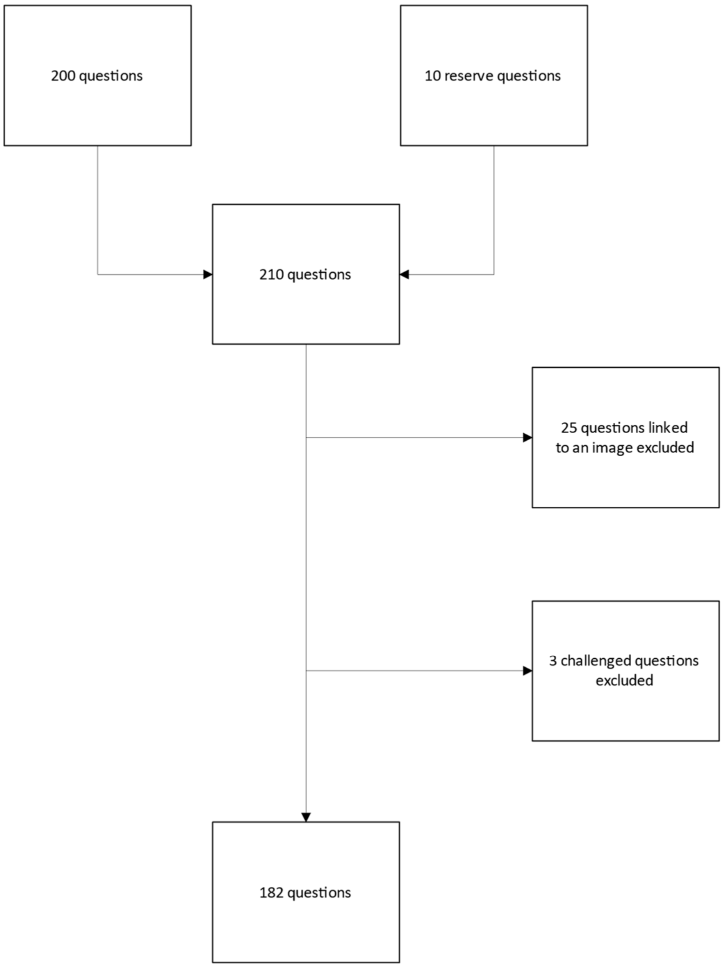

200 questions
10 reserve questions
210 questions
25 questions linked to an image excluded
3 challenged questions excluded
182 questions

Figure 1. Study Flow Diagram: Question Selection Process and Exclusion Criteria.

We prompted the questions to the GPT4 (August 2023 version) and to GPT-3.5 [18]. We prompted the questions to ChatGPT in Spanish and English. We used the prompt: "*Please answer the following multiple-choice questions. Note: These questions are from the Medical Intern Resident (MIR) exam taken by doctors in Spain. The answers to the questions are universal; however, some questions may have nuances specific to Spain, especially in areas related to the vaccination calendar, list of notifiable diseases, legal aspects, and organization of health services. Answer with only the question number and the number of the correct answer option*". We prompted the questions in blocks of 10 questions.

Questions were classified into two groups: Theoretical and Practical. Questions were also classified according to the specialty in the following groups: Cardiology, Dermatology, Endocrinology, Epidemiology, Ethics, Family and Community Medicine, Gastroenterology, Genetics, Geriatrics, Hematology, Immunology, Infectious Diseases, Legal and Forensic Medicine, Maxillofacial Surgery, Nephrology, Obstetrics and Gynecology, Orthopedic and Traumatology Surgery, Otorhinolaryngology, Pathology, Pharmacology, Physiopathology, Plastic Surgery.

The English translation of the questions and the correct answers are presented in Supplements S1 and S2. We computed the number of words and characters for each question in Spanish using Microsoft Word 365 (Microsoft® Word 365 MSO version 2307). We also calculated the number of tokens using the AI Tokenizer (https://platform.openai.com/tokenizer).

To test the existence of nonrandom error, we submitted the test 3 times to Chat GPT4, two times with the questions in the original order and another time with the questions in a random order, and comparisons were made between the different versions.

We assessed the potential risks to the patient in a hypothetical case scenario in which the failure to adequately answer the question would have occurred in real life. We evaluated the potential risks for patients of failing key questions using the National Coordinating Council for Medication Error Reporting and Prevention (NCC MERP) classification system [19]. This Classification has four categories of error: no error, error-no harm, error- harm and error-death. There are also the following subcategories:

- Category A. Circumstances or events that have the capacity to cause error.
- Category B. An error occurred, but the error did not reach the patient.
- Category C. An error occurred that reached the patient but did not cause the patient harm.
- Category D. An error occurred that reached the patient and required monitoring to confirm that it resulted in no harm to the patient or required intervention to preclude harm.
- Category E. An error that may have contributed to or resulted in temporary harm to the patient and required intervention.
- Category F. An error occurred that may have contributed to or resulted in temporary harm to the patient and required initial or prolonged hospitalization.
- Category G. An error occurred that may have contributed to or resulted in permanent patient harm.
- Category H. An error occurred that required intervention necessary to sustain life.
- Category I. An error occurred that may have contributed to or resulted in the patient's death. [19].

The questionnaire in Spanish was presented twice to the GPT-4 in its original sequence, termed "Original Question Sequence," to discern any patterns in the responses. These attempts were called the "First Attempt" and "Second Attempt," respectively. Subsequently, the questions were reordered randomly to form the "Random Question Sequence." GPT-4's responses to this shuffled set were then evaluated in two distinct manners: once in the randomized sequence itself ("Evaluated in Random Order") and once after sorting the outcomes according to the original MIR examination sequence ("Evaluated in Original Sequence"). The consistency and predictability of the answers across these scenarios were quantified using the Runs Test, with various metrics such as the Test Value, Number of Runs, Z-value, and Asymptotic Significance being recorded for a comprehensive comparison.

2.3. Statistical Analysis

We computed confidence intervals of proportions and Cohen's Kappa, a statistic used to measure inter-rater reliability for categorical items. The calculations were made with Openepi [20]. All the remaining analyses were performed with IBM SPSS version 26. We calculated the Wald-Wolfowitz Runs Test, a non-parametric statistical test that checks the randomness of a data sequence to see if there was an aggregation on the error. We compared the results using the McNemar Test between the GPT-3.5 and GPT-4 versions and the Spanish and English versions of the questionnaire. We computed Cohen's Kappa to study the concordance of the results between the first and second attempts to submit the examinations to GPT-4. We also calculated the Chi-square test, as well as adjusted standardized residuals.

We used logistic multivariate regression models to see if the number of words, characters, and tokens were related to failing a question. We also used Polynomial Logistic Regression, introducing square terms in the models. As polynomial terms, especially when squared or cubed, can lead to multi-collinearity problems, we centered the variables (subtracting their mean) before squaring.

We also performed logistic regression using lags (lag1 and lag2) of the correct question with the results of both the Spanish version of the questionnaire and the random version with GPT-4.

3. Results

The LLM GPT-4 proportion of correct answers in Spanish was 86.81%, higher than that of GPT-3.5 at 63.18%. ($p< 0.001$) (Table 1) The same happened with the English translation of the questionnaire. When the English questionnaire was used, responses improved slightly by 1.10% with GPT-4 and 3.30% with GPT -3.5, but these differences were not statistically significant.

Table 1. Comparison of Correct Response Proportions for Input in Spanish and English: GPT-3.5 vs. GPT-4, with McNemar Test Results.

Language	GPT-4 % (95% CI)	GPT-3.5 % (95% CI)	Sig**
Spanish	86.81 (81.13-90.98)	63.18 (55.98-69.85)	< 0.001
English	87.91 (82.38-91.88)	66.48 (59.35-72.94)	< 0.001
Sig*	0.824	0.441	N=182

N=182 *comparison between languages within versions **Comparison between versions within languages.

From here on, we will only report on GPT-4 with questions in Spanish. There was no difference in the results between theoretical and practical questions in the examination (Table 2).

Table 2. Comparison of Correct Response Proportions for MIR Examination Theoretical and Practical Questions.

	GPT-4 % correct answers (N=182)	Significance†
Type of questions		0.927
Theoretical	87.1% (85)	
Practical	86.6% (97)	

N=Total of questions † Chi square.

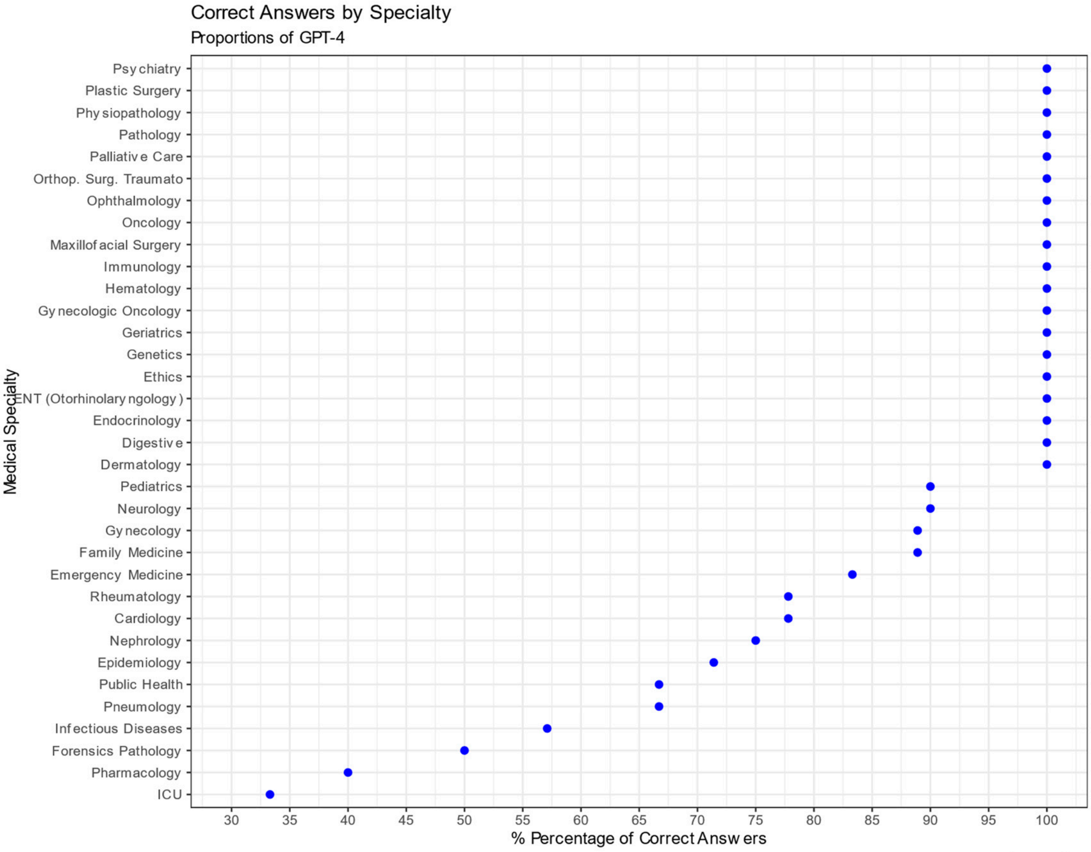
The analysis of GPT-4's performance on the Spanish MIR examination for physician residency program entrance across different medical specialties is presented in Table 3 and Figure 2. The total number of questions evaluated across all specialties was 182. GPT-4 exhibited varied success rates among the specialties. Specialties such as Pathology, Orthopedic Surgery and Traumatology, Dermatology, Digestive, Endocrinology, Plastic Surgery, Ethics, Physiopathology, Genetics, Geriatrics, Hematology, Immunology, Maxillofacial Surgery, Ophthalmology, Gynecologic Oncology Oncology, ENT (Otorhinolaryngology), and Psychiatry yielded a 100% correct response rate. However, there were areas where GPT-4's performance was significantly below expectations based on the adjusted residual analysis, such as in Pharmacology (40%, p<0.001), Critical Care medicine (33.3%, p<0.01), and Infectious Diseases (57.1%, p<0.05).

Table 3. Comparison Correct Response Proportions of GPT-4 for MIR Examination in Spanish by Specialty.

Specialty	% Correct	N	Specialty	% Correct	N
Pathology	100%	1	Forensics & legal medicine	50.0%	2
Cardiology	77.8%	9	Family Medicine	88.9%	9
Orthopedic Surgery and Traumatology	100%	10	Nephrology	75.0%	4
dermatology	100%	1	Pneumology	66.7%	6
digestive	100%	6	Neurology	90.0%	10
endocrinology	100%	8	ophthalmology	100%	3
epidemiology	71.4%	7	gynecologic oncology	100%	2
Plastic Surgery	100%	2	Oncology	100%	2
Ethics	100%	4	ENT (Otorhinolaryngology)	100%	3
pharmacology	40%***	5	palliative care	100%º	2
physiopathology	100%	7	Pediatrics	90.0%	10

genetics	100%	2	Public Health	66.7%	3
geriatrics	100%	6	Psychiatry	100%	8
gynecology	88.9%	9	Rheumatology	77.8%	9
hematology	100%	5	Critical Care	33.3%**	3
infectious diseases	57.1*	7	emergency medicine	83.3%	6
immunology	100%	7			
Maxillofacial surgery	100%	2	TOTAL		182

Adjusted residual analysis *p <0.05 **p <0.01 *** p < 0.001.



Source: Own

Figure 2. Proportions of GPT-4 correct answers by specialty.

We observed consistency in evaluating GPT-4's performance on the Spanish MIR examination across two submissions. The first attempt resulted in an 86.8% accuracy rate, while the second attempt showed a slight increase, achieving an accuracy of 89%. The agreement between the two submissions, as measured by Cohen's Kappa interrater, was substantial, with a value of 0.7418 (95% CI: 0.5973 - 0.8863), suggesting consistent performance across both trials. Moreover, the McNemar's test revealed no statistically significant differences between the two attempts (p = 0.344).

Table 4. Comparison of Correct Response Proportions for first and second attempts in Spanish using GPT-4, with McNemar Test Results.

	Second Attempt		
First Attempt	Wrong	Correct	Total
Wrong	17	7	24
Correct	3	155	158
Total	20	162	182

McNemar Test $P_{(exact)}$=0.344 Cohen's Kappa, = 0.74 (95% CI: 0.59 - 0.88).

Table 5 showcases GPT-4's responses to the MIR Examination questions across varied sequence scenarios. In the original order, differences between the two attempts were notable in the number of wrongly presented values of 24 and 20. The first and second attempts demonstrated exact significances of 0.017 and 0.032, respectively, indicating statistically significant nonrandom patterns. When evaluating the randomized questions using the original sequence, a significance of 0.040 was observed, suggesting a similar nonrandom trend. However, a clear shift occurred when the randomized responses were assessed in their inherent order: the significance reached 1, pointing to a more randomized response pattern in this context. Considering the results, it is evident that GPT-4's responses to the MIR Examination's questions manifest nonrandom patterns in both original sequence attempts and when the randomized sequence was evaluated against the original order. However, the model exhibits a trend towards randomness when the shuffled questions were assessed in their original order, highlighting the crucial role of question sequencing in dictating AI response tendencies.

Table 5. Comparison of GPT-4's Responses to MIR Examination Questions: Effects of Question Order on Runs Test Outcomes.

	Original Question Sequence		Random Question Sequence	
Test Scenario	**1st Attempt**	**2nd attempt**	**Evaluated with the Original Sequence**	**Evaluated with the Random Order**
Test Value Median	1	1	1	1
Wrong answers	24	20	23	23
Correct answers	158	162	159	159
Total Questions	182	182	182	182
Number of Runs	35	31	35	41
Z	-2.507	-2.148	-2.097	-0.063
Exact Significance (2-tailed)	0.017	0.032	0.040	1.000

We did a univariate logistic regression and multivariate logistic regression to determine if there was an association between the length of the questions and the success in answering the questions. The results are presented in Table 6. The number of words, characters, and tokens did not influence the examination performance of GPT-4.

Table 6. Univariate and Multivariate logistic regression of GPT-4's Responses to MIR Examination Questions: Effects of number of words, characters, and tokens (all in hundreds).

	Univariate logistic regression		Multivariate logistic regression	
Length of the question	**OR (95% CI)**	**p**	**OR (95% CI)**	**p**
Number of Words*	1.82 (95% CI 0.51-6.47)	0.351	1.33 (95% CI 0.03-53.57)	0.880
Number of Characters*	1.09 (95% CI 0.90-1.34)	0.370	1.70 (95% CI 0.01-259.69)	0.835
Number of Tokens*	1.30 (95% CI 0.74-2.33)	0.361	0.92 (95% CI 0.24-3.45)	0.896

* In hundreds.

We computed a Polynomial Logistic. The results suggest that there is no statistically significant association between the length of the MIR Examination questions (measured in hundreds of words, characters, or tokens) and the success rate of GPT4 in answering them, both in terms of linear and quadratic effects, (Table 7) indicating that the question length in the measured units does not significantly influence GPT4's performance on the MIR test.

Table 7. Univariate Polynomial Logistic Regression of GPT-4's Responses to MIR Examination Questions: Effects of number of words, characters, and tokens (all in hundreds).

	Univariate Polynomial Logistic Regression	
Length of the question	**OR (95% CI)**	**p**
Words		
words	1.86 (0.51-6.80)	0.344
$words^2$	0.84 (0.06-11.10)	0.893
Characters		
Characters	1.11 (0.91-1.36)	0.293
$Characters^2$	0.983(0.92-1.05)	0.588
Tokens		
Tokens	1.38 (0.77-2.50)	0.283
$Tokens^2$	0.88 (0.55-1.41)	0.597

Our analysis of the 182 questions from the Spanish MIR examination using GPT4 resulted in an error rate of 13.2%, with 24 questions answered incorrectly. Delving into the potential implications of these errors, as described in Table 8 using the NCC MERP Classification System, the "No error" rate was 5.5%. In contrast, the "Error no harm" category and the "Error harm" category, suggesting direct harm to patients if such errors occurred in real-world medical scenarios, resulted in a rate of 3.3%.The complete list of questions failed by GPT-4, with its answer and the correct answer is found in Table S3

Table 8. Distribution and Potential Risk of Errors Identified using the categories of the NCC MERP Classification System.

Type of Error	n	%	Rate%(95% CI)
1. No error	10	41.7	5.5 (3.0-9.8.)
2. Error no harm	8	33.3	4.4 (2.2-8.4)
3. Error harm	6	25.0	3.3 (1.5-7.0)
Total Incorrect Answers	24	100	
Correct Answers	158		
Total Questions	182		

Table 9 provides more granulated insights into the nature of these errors using the NCC MERP subcategories. The predominant error type was "A. capacity to cause an error," which occurred in 41.7% of the errors with a rate of 5.5%. It is crucial to highlight the absence of errors in the gravest categories, "H. error required intervention to sustain life" and "I. error contributed to or resulted in death," showing a 0% rate. This analysis underscores the importance of understanding the quantity of errors and their qualitative severity when assessing potential patient risks when failing to correctly answer the MIR questions.

Table 9. Distribution and Potential Risk of Errors Identified using the NCC MERP Classification System subcategories.

Type of Error	n	%	Rate% (95% CI)
- A. Capacity to cause error.	10	41.7	5.5 (3.0-9.8)
- B. Error did not reach the patient.	1	4.4	5.4 (0.9-3.0)
- C. Error reached patient did not cause harm.	3	12.5	1.6 (0.6-4.7)
- D. Error reached the patient and required monitoring	4	16.7	2.2 (0.9-5.5)
- E. Error temporary harm and required intervention.	2	8.3	1.1 (0.3-3.9)
- F. Error required hospitalization.	2	8.3	1.1 (0.3-3.9)
- G. Error resulted in permanent patient harm.	2	8.3	1.1 (0.3-3.9)
- H. Error required intervention to sustain life.	0	0	0 (0-2.0.)

- I. Error contributed to or resulted in the death	0	0	0 (0-2.0.)
Total Incorrect Answers	24	100	
Correct Answers	158		
Total Questions	182		

We found an association between the Potential Risk of Errors and specialty (p= 0.01). All the errors in Cardiology and the Critical Care produced Harm (p < 0.01), while the errors in Pneumology did not produce harm.(p <0.05) (Table 10)

Table 10. Distribution of Potential Risk of Errors Identified using the NCC MERP Classification by specialty and System categories.

Specialty	No error n(%)	No harm n(%)	Harm n(%)	Total
Cardiology	0	0	2** (100%)	2 (100%)
Epidemiology	2(100%)	0	0	2 (100%)
Pharmacology	0	2 (66.6%)	1	3 (100%)
Gynecology	0	1 (100%)	0	1 (100%)
Infectious Diseases	0	2 (66.6%)	1 (33.3%)	3 (100%)
Forensics & Legal Medicine	1 (100%)	0	0	1 (100%)
Family Medicine	1 (100%)	0	0	1 (100%)
Nephrology	1 (100%)	0	0	1 (100%)
Pneumology	0	2* (100%)	0	2 (100%)
Neurology	1 (100%)	0	0	1 (100%)
Pediatrics	1 (100%)	0	0	1 (100%)
Public Health	0	1 (100%)	0	1 (100%)
Rheumatology	2 (100%)	0	0	2 (100%)
Critical Care	0	0	2** (100%)	2 (100%)
Emergency medicine	1 (100%)	0	0	1 (100%)
TOTAL	10(41.7%)	8 (33.3%)	6 (25%)	24 (100%)

X^2= 38.667 , df=28, $P_{(exact)}$=0.011 Adjusted Standardized Residuals * p < 0.05 **p < 0.01.

4. Discussion

We analyzed the comparative performance between GPT-3.5 and GPT-4 in handling the Spanish MIR examination and provide essential insights into the advancements ofLLM in medical knowledge comprehension and accuracy. The improved consistency demonstrated by GPT-4 across multiple attempts demonstrates the refinements in training and underlying model architecture in the GPT-4 version. This superiority of GPT-4 over GPT-3.5 has also been shown in drug information queries [21,22].

Although specialties such as Surgery and Pediatrics saw commendable success rates, GPT – has a lower performance in Pharmacology, Critical Care, and Infectious Diseases, highlighting that there might be areas of medical knowledge that require a more nuanced understanding or perhaps more excellent representation in training data.

In the same way, a study in which GPT-4 was presented with ten antidepressant-prescribing vignettes found that the model seems to recognize and utilize several standard strategies often used in psychopharmacology. However, due to some recommendations that may not be ideal, relying on LLM for psychopharmacologic treatment guidance without additional oversight remains risky [23].

The length of the question had no significant bearing on GPT-4's performance, which challenges our hypothesis that lengthier inputs might pose comprehension issues for the model.

The rate of success of GPT-4 (86.8%) was very close to the physician who scored the highest on the MIR examination (91.5%) [24] The number of correct answers by GPT could be higher because ten more questions were challenged, although the Ministry of Health did not accept them. The correct

doi:10.20944/preprints202309.1272.v1

answer [25], of one of the challenged questions coincided with the one chosen by ChatGPT. In addition, ChatGPT failed in two questions that the academies considered unchallengeable.

A previous Spanish study submitted the same MIR examination to GPT3.5 and obtained a lower score (54.8%) than ours [26]. One explanation of the difference could be the prompt used. In our study, we situated the questions in a context, while the other study asked the model to fill in the questions. How the prompt is formulated may affect the results.

Though relatively low, the GPT-4's error rate of 13.2% should be a concern, especially when contextualized within the medical field where stakes are incredibly high. The categorization of these errors through the NCC MERP Classification System reveals that while most of the errors might be benign, the existence of even a small percentage that could lead to patient harm emphasizes the importance of human oversight. Such oversight is an ethical obligation and could serve as a mitigation strategy in preventing the potential pitfalls of over relying on AI recommendations without expert verification.

We posit that while GPT-4 and similar models represent significant advancements in AI-driven knowledge databases and comprehension, their function should remain supportive, especially in sensitive sectors like Medicine. Medical practitioners must continue to utilize their training, experience, and intuition, using AI tools as complementary resources. In a study evaluating ChatGPT's clinical decision support using standardized clinical vignettes, the model achieved a 71.7% overall accuracy, excelling in final diagnosis tasks (76.9%) but showing lower performance in the initial differential diagnosis (60.3%) [27]. Another study on triage with clinical vignettes in ophthalmology found almost the same success rate as ophthalmologists in training, 93% of the model versus 95% of the residents. The same happened in predicting the correct diagnosis of common urinary diseases. ChatGPT had a higher accuracy rate (91%) in predicting the proper diagnosis of common urinary diseases than junior urology residents [28]. Chat GPT also has shown the ability to evaluate neuro-exams using established assessment scales [29]. It has outperformed medical students and neurosurgery residents on neurosurgery test examinations with a 79% success [30].

Another potential use of ChatGPT could be to answer patient questions. ChatGPT has been tested with real questions that patients ask physicians. The average accuracy scores (1 to 5) for questions about treatments, symptoms, and diagnostic tests were 3.9, 3.4, and 3.7, respectively [8]. GPT could be used in health education to help patients understand their diseases, treatment, and potential complications [31]and adapt their lifestyle to conditions such as metabolic syndrome or obesity [32,33].

Another use of ChatGPT could be to help physicians generate reports, make them comprehendible for patients [34] or even write them in another language. This has been shown especially in the field of Radiology [35,36]. The sequencing of the MIR Examination questions is crucial in understanding GPT-4's performance. Table 3 demonstrates that in evaluating GPT-4's responses, it becomes clear that questions, when grouped by specialty, can create distinct patterns in the model's response. For instance, GPT-4 managed notable success rates in the areas of Neurology (90.0%) and Cardiology (77.8%), whereas it grappled with specialties such as Pharmacology, achieving only a 40% success rate.

This specialty-based pattern becomes even more significant when the examination's questions are shuffled, as observed in Table 4. When the questions were randomized, GPT-4's performance, as measured by the Runs Test, changed. The model showed a nonrandom response pattern during the first three tests, with exact significance levels of 0.017, 0.042, and 0.040, suggesting that the model had difficulty with specific clusters of questions. However, when the randomized question order was evaluated as per the random sequence, the significance value rose sharply to 1.000, indicating a shift to a more random response pattern. This data implies that what we initially thought might be 'model fatigue' is more about the organization and sequence of the examination's design. If the model found a series of questions challenging, it is likely because those questions were from a particularly tough specialty, not because of the sequence. This pattern can be misleading and interpreted as fatigue, especially when questions of a challenging specialty are clustered together. Our analysis underlines the need to consider not only the nature of the questions but also their internal organization and sequencing when evaluating AI performance, particularly with high stakes exams. While intriguing, the notion of 'model fatigue' does not hold water when faced with the intricacies of the examination structure and the model's response patterns.

4.1. LLM variable responses

We found variability in GPT-4's responses. Even when identical prompts are used, there can be some variability. This variability is an intriguing aspect of LLM design and offers a glimpse into the complexities of neural networks. Unlike deterministic systems (like traditional statistical packages), which consistently produce unchanging outputs for a given input, LLM like GPT-4 operate in a probabilistic domain. The model identifies plausible outcomes for any given information and selects the most probable one. However, what is deemed "most probable" can vary significantly with each invocation due to the model's stateless nature and inherent randomness.

Additionally, external factors like degradation in performance, technical glitches, or even subtle model updates can further introduce variability. The prompt's ambiguity can also lead the model to be interpreted differently in separate instances. Though these variations might be seen as inconsistencies, they are a feature of the model's dynamic and probabilistic architecture, setting it apart from traditional deterministic systems.

In the design phase, we had to decide between submitting the questions individually and raising the questions in blocks of 10. We choose to submit in 10-question blocks. This approach is faster and more efficient, especially when dealing with many questions. In addition, GPT-4 in August 2023 had a limitation of a maximum of 50 questions every 3 hours. This posed several inconveniences: GPT-4 has a token limit (the exact number can depend on the specific version/configuration of the model). The model will not "see" the entire block if a block exceeds this limit. There could be Potential Carry-over Effects: The context from one question could unintentionally influence the model's response to a subsequent question, which might skew results if not desired. Submitting the questions one by one could have the advantage of treating each question in isolation, eliminating potential contextual influence from previous questions, and allowing a more precise assessment of GPT-4's performance on individual questions. It also ensures that each question is processed with the same amount of attention without being influenced by the "length" or complexity of prior questions in a session.

On the other hand, GPT-4 does not carry over any information from the previous question, which can be desirable if we want to ensure no "memory" effects between questions. Nevertheless, submitting one question isolated every time is more time-intensive than offering them in batches. Future research could test both strategies by submitting in batches or one by one and determine the best number of questions in the batches.

4.2. Limitations of the study

Our study has several limitations. First, the number of questions in each specialty is tiny, so the confidence intervals for GPT success are enormous. Furthermore, in the future, evaluation of more tests to estimate the success rate more accurately is needed. It would also be interesting to ask the model for the reasons for choosing the wrong answer to determine the causes of the failures. Progressively, it could be tested as a diagnostic aid in those specialties with higher success rates.

Second, LLM models can show highly variable accuracy across different domains as our study results demonstrates. Our study corroborates earlier observations on the performance of GPT-4 in the MIR examination across different medical specialties and shows promise when applied in various medical disciplines such as allergology, dermatology, and radiology [1–3]. GPT-4 achieved a perfect score in specialties including Dermatology, aligning with prior literature highlighting its efficacy in that field. Amazingly, in our study, GPT-4 chose the correct answer in 100% of the Gastroenterology (digestive) questions, but another study failed the Gastroenterology examination, not reaching the score of 70% necessary to pass the test [7]. The variability in performance across specialties does support our initial hypothesis, indicating that while GPT-4 shows exemplary results in some areas, it has potential weaknesses in others, such as Pharmacology and Critical Care. This differential performance suggests that navigating specialized medical examinations like the MIR demands a broader and more nuanced knowledge base. It is crucial to understand where AI tools like ChatGPT excel, and how they can be best complement human expertise.

Third, the obsolescence index applies to LLM searches and measures how quickly information becomes outdated. For instance, the obsolescence index for medical textbooks of Internal Medicine is around five years, meaning that a fresh out-of-print book has five years of information. On the other hand, with the time considered for writing, peer review, and journal submission, scientific papers may have an obsolescence index of one year [37,38]. This rapid turnover in medical knowledge underscores the importance of regular updates to AI models to ensure their recommendations remain current and clinically relevant. A testament to the vast and ever-growing body of medical information is a recent search conducted by the authors on PubMed, on 19/08/2023, which yielded a staggering 387,974 papers related to COVID-19 alone. With search terms like "coronavirus," "covid," "covid-19," and "sars-cov-2," this immense volume of information is impossible for any individual physician to sift through comprehensively. However, this is precisely where artificial intelligence systems can excel. By rapidly processing and analyzing vast datasets, AI can assist in synthesizing pertinent information and insights, ensuring that medical professionals can be assisted with the most up-to-date knowledge. However, a judicious approach is essential; while AI can provide quick answers, physicians must remain vigilant, cross-referencing recommendations, especially in areas where the AI's performance has shown high variability and being mindful of potential "hallucinations" or inaccuracies in AI outputs. A hallucination is a factual inaccuracy that appears to be scientifically plausible [39] [40] [41] but is a fabrication [42]. GPT also can generate fake bibliographic references [43] [44]. For example, one study about kidney transplantation found that ChatGPT failed to provide references for any of the scientific data it provided regarding kidney transplants, and when requested for references, it provided inaccurate ones [45]. Due to that, it is always essential to verify the information provided by ChatGPT [46].

Fourth, LLM have yet been shown to evaluate or improve hospital-specific activities such as medical documentation, e.g., coding causes of hospitalization or death with the International Classification of Diseases, infection control, and reviewing medical records to detect whether a patient has had a healthcare-associated infection. In the continually evolving landscape of medicine, the potential applications of advanced tools like GPT in clinical practice are manifold. ChatGPT is adept at handling real-world patient queries [4]. Patients can also use chat GPT to seek information when they fear sharing their doubts on sensible issues with their physicians [47]. GPT, first unveiled September 2021, offers a myriad of uses, from aiding physicians in differential diagnosis and treatment recommendations to facilitating patient education and streamlining telemedicine consultations. It can also assist medical research, providing rapid summaries of the latest studies or advancements in specific domains.

Fifth, first of all, the study focused solely on the Spanish MIR Medical Examination which may limit the generalizability of the findings to other medical examinations or languages.

4.3. Implications of This Study

The prospects of integrating AI tools like GPT into medicine are promising, and a collaborative approach remains paramount, where AI can be sued to complement rather than endeavor to replace human expertise. Future research should explore mechanisms to integrate AI feedback with human expertise, potentially through hybrid decision-making systems, to harness the strengths of both systems. ChatGPT could be used by physicians as a diagnostic assistance tool, helping physicians with differential diagnoses [2], and also help personalize treatments based on existing evidence and the scientific literature [2]. As technology continues to evolve and refined models emerge, continuous evaluation, as was done in this study, will be critical. Such evaluations ensure that while we tap into the immense potential of AI, we remain grounded in the core objective: enhancing patient care and medical outcomes. We need to recognize that the latest advances in medical and computer technology, however, are raising many questions relative to ethics of real-world care and how and pledge to support ethical medical technologies with the goal of *"Primum non nocere."* [48]

5. Conclusions

This study illuminates the advanced performance of LLM and the differences between GPT-3.5 and GPT-4 in addressing the Spanish MIR examination for medical specialization. Significantly, GPT-4 showcased a superior accuracy rate in both Spanish and English, with consistency evident across multiple submissions. Despite the impressive success rates across various specialties, some areas like Pharmacology, Critical Care, and Infectious Diseases revealed notable shortcomings. Interestingly, when quantified in terms of words, characters, or tokens, the question length did not influence GPT-4's performance. While the model's error rate stood at 13.2%, the potential implications of these mistakes, when framed using the NCC MERP Classification System, pointed towards a low percentage of errors that could result in direct harm. AI models like GPT-4 exhibit strong capabilities, but their application in medicine requires meticulous scrutiny and emphasizes the need for further improvements in LLMs before they can be reliably deployed in medical settings for decision-making support.

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on Preprints.org, Table S1: English Translation of the MIR Examination; Table S2: Answer to MIR examination questions., Table S3: Questions failed by GPT4 in the MIR examination, indicating the correct answer, the wrong answer, and the Error classification.

Author Contributions: Conceptualization FGG EAO IAO; methodology FGG, SGA, LGA, RAB, LO, EAO, PB, IAO; software FGG SGA LGA RAB LO; investigation SGA LGA RAB LO RM; formal analysis RAB, WO, RM, EAO, PB, IAO; writing—original draft preparation, FGG, SGA, LGA, RAB, LO, WO, RM, EAO, PB, IAO; writing—review and editing FGG, SGA, LGA, RAB, LO, WO, RM, EAO, PB, IAO; visualization SGA, LGA, WO; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data base can be obtained from the authors on reasonable request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Sohail, S. S. A Promising Start and Not a Panacea: ChatGPT's Early Impact and Potential in Medical Science and Biomedical Engineering Research. *Ann Biomed Eng* **2023**. https://doi.org/10.1007/s10439-023-03335-6.
2. Goktas, P.; Karakaya, G.; Kalyoncu, A. F.; Damadoglu, E. Artificial Intelligence Chatbots in Allergy and Immunology Practice: Where Have We Been and Where Are We Going? *J Allergy Clin Immunol Pract* **2023**. https://doi.org/10.1016/j.jaip.2023.05.042.
3. Wiens, J.; Mihalcea, R.; Nallamothu, B. K. Current Large Language Models Will Not Fix Health Care. Here's What Could. *Stat* **2023**, No. Aug. 25, 2023.
4. Dunn, C.; Hunter, J.; Steffes, W.; Whitney, Z.; Foss, M.; Mammino, J.; Leavitt, A.; Hawkins, S. D.; Dane, A.; Yungmann, M.; Nathoo, R. Artificial Intelligence–Derived Dermatology Case Reports Are Indistinguishable from Those Written by Humans: A Single-Blinded Observer Study. *J Am Acad Dermatol* **2023**. https://doi.org/10.1016/j.jaad.2023.04.005.
5. Shen, Y.; Heacock, L.; Elias, J.; Hentel, K. D.; Reig, B.; Shih, G.; Moy, L. ChatGPT and Other Large Language Models Are Double-Edged Swords. *Radiology* **2023**, *307* (2). https://doi.org/10.1148/radiol.230163.
6. Johnson, D.; Goodman, R.; Patrinely, J.; Stone, C.; Zimmerman, E.; Donald, R.; Chang, S.; Berkowitz, S.; Finn, A.; Jahangir, E.; Scoville, E.; Reese, T.; Friedman, D.; Bastarache, J.; van der Heijden, Y.; Wright, J.; Carter, N.; Alexander, M.; Choe, J.; Chastain, C.; Zic, J.; Horst, S.; Turker, I.; Agarwal, R.; Osmundson, E.; Idrees, K.; Kiernan, C.; Padmanabhan, C.; Bailey, C.; Schlegel, C.; Chambless, L.; Gibson, M.; Osterman, T.; Wheless, L. Assessing the Accuracy and Reliability of AI-Generated Medical Responses: An Evaluation of the Chat-GPT Model. *Res Sq* **2023**. https://doi.org/10.21203/rs.3.rs-2566942/v1.
7. Suchman, K.; Garg, S.; Trindade, A. J. Chat Generative Pretrained Transformer Fails the Multiple-Choice American College of Gastroenterology Self-Assessment Test. *American Journal of Gastroenterology* **2023**, *Publish Ah*. https://doi.org/10.14309/ajg.0000000000002320.

8. Lahat, A.; Shachar, E.; Avidan, B.; Glicksberg, B.; Klang, E. Evaluating the Utility of a Large Language Model in Answering Common Patients' Gastrointestinal Health-Related Questions: Are We There Yet? *Diagnostics* **2023**, *13* (11), 1950. https://doi.org/10.3390/diagnostics13111950.
9. Gilson, A.; Safranek, C. W.; Huang, T.; Socrates, V.; Chi, L.; Taylor, R. A.; Chartash, D. 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. https://doi.org/10.2196/45312.
10. Epstein, R. H.; Dexter, F. Variability in Large Language Models' Responses to Medical Licensing and Certification Examinations. Comment on "How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education a. *JMIR Med Educ* **2023**, *9*, e48305. https://doi.org/10.2196/48305.
11. Jung, L. B.; Gudera, J. A.; Wiegand, T. L. T.; Allmendinger, S.; Dimitriadis, K.; Koerte, I. K. ChatGPT Passes German State Examination in Medicine with Picture Questions Omitted. *Dtsch Arztebl Int* **2023**. https://doi.org/10.3238/arztebl.m2023.0113.
12. Wang, H.; Wu, W.; Dou, Z.; He, L.; Yang, L. Performance and Exploration of ChatGPT in Medical Examination, Records and Education in Chinese: Pave the Way for Medical AI. *Int J Med Inform* **2023**, *177*, 105173. https://doi.org/10.1016/j.ijmedinf.2023.105173.
13. Kao, Y.-S.; Chuang, W.-K.; Yang, J. Use of ChatGPT on Taiwan's Examination for Medical Doctors. *Ann Biomed Eng* **2023**. https://doi.org/10.1007/s10439-023-03308-9.
14. Takagi, S.; Watari, T.; Erabi, A.; Sakaguchi, K. Performance of GPT-3.5 and GPT-4 on the Japanese Medical Licensing Examination: Comparison Study. *JMIR Med Educ* **2023**, *9*, e48002. https://doi.org/10.2196/48002.
15. Levin, G.; Horesh, N.; Brezinov, Y.; Meyer, R. Performance of ChatGPT in Medical Examinations: A Systematic Review and a Meta-analysis. *BJOG* **2023**. https://doi.org/10.1111/1471-0528.17641.
16. Orden SND/840/2022, de 26 de Agosto, Por La Que Se Aprueba La Oferta de Plazas y La Convocatoria de Pruebas Selectivas 2022 Para El Acceso En El Año 2023, a Plazas de Formación Sanitaria Especializada Para Las Titulaciones Universitarias de Grado/Licencia. *Boletin Oficial del Estado* **2022**, No. 211, 122047–122309.
17. Gamarra, M. Resultados de los extracomunitarios en el MIR 2023 https://www.consalud.es/especial-mir/mir-2023-46-plazas-han-ido-parar-extranjeros_129841_102.html (accessed 2023 -05 -11).
18. OpenAI. GPT-4 Technical Report. **2023**.
19. NCC-MERP. National Coordinating Council for Medication Error Reporting and Prevention.Taxonomy of Medication Errors https://www.nccmerp.org/sites/default/files/taxonomy2001-07-31.pdf (accessed 2023 -08 -15).
20. Dean AG, Sullivan KM, S. MM. OpenEpi: Open Source Epidemiologic Statistics for Public Health, Versión. 2013.
21. He, N.; Yan, Y.; Wu, Z.; Cheng, Y.; Liu, F.; Li, X.; Zhai, S. Chat GPT-4 Significantly Surpasses GPT-3.5 in Drug Information Queries. *J Telemed Telecare* **2023**. https://doi.org/10.1177/1357633X231181922.
22. Kleebayoon, A.; Wiwanitkit, V. Correspondence on Chat GPT-4, GPT-3.5 and Drug Information Queries. *J Telemed Telecare* **2023**. https://doi.org/10.1177/1357633X231189760.
23. Perlis, R. H. Research Letter: Application of GPT-4 to Select next-Step Antidepressant Treatment in Major Depression. *medRxiv* **2023**. https://doi.org/10.1101/2023.04.14.23288595.
24. Galvan, A. Patricia Andrés, número 1 del examen MIR 2023, elige Dermatología para realizar su Residencia https://aedv.es/patricia-andres-elige-dermatologia-para-realizar-residencia/#:~:text=Natural de Bilbao y alumna,y 116%2C9836 puntos totales.
25. Examen MIR 2023: ¿Qué preguntas podrían ser impugnables? https://www.diariomedico.com/medicina/medico-joven/mir/examen-mir-2023-que-preguntas-podrian-ser-impugnables.html.
26. Carrasco, J. P.; García, E.; Sánchez, D. A.; Porter, E.; De La Puente, L.; Navarro, J.; Cerame, A. ¿Es Capaz "ChatGPT" de Aprobar El Examen MIR de 2022? Implicaciones de La Inteligencia Artificial En La Educación Médica En España. *Revista Española de Educación Médica* **2023**, *4* (1). https://doi.org/10.6018/edumed.556511.
27. Rao, A.; Pang, M.; Kim, J.; Kamineni, M.; Lie, W.; Prasad, A. K.; Landman, A.; Dreyer, K.; Succi, M. D. Assessing the Utility of ChatGPT Throughout the Entire Clinical Workflow: Development and Usability Study. *J Med Internet Res* **2023**, *25*, e48659. https://doi.org/10.2196/48659.
28. Xv, Y.; Peng, C.; Wei, Z.; Liao, F.; Xiao, M. Can Chat-GPT a Substitute for Urological Resident Physician in Diagnosing Diseases?: A Preliminary Conclusion from an Exploratory Investigation. *World J Urol* **2023**. https://doi.org/10.1007/s00345-023-04539-0.
29. Chen, T. C.; Kaminski, E.; Koduri, L.; Singer, A.; Singer, J.; Couldwell, M.; Delashaw, J.; Dumont, A.; Wang, A. Chat GPT as a Neuro-Score Calculator: Analysis of a Large Language Model's Performance on Various Neurological Exam Grading Scales. *World Neurosurg* **2023**. https://doi.org/10.1016/j.wneu.2023.08.088.

30. Guerra, G.; Hofmann, H.; Sobhani, S.; Hofmann, G.; Gomez, D.; Soroudi, D.; Hopkins, B. S.; Dallas, J.; Pangal, D.; Cheok, S.; Nguyen, V.; Mack, W. J.; Zada, G. GPT-4 Artificial Intelligence Model Outperforms ChatGPT, Medical Students, and Neurosurgery Residents on Neurosurgery Written Board-like Questions. *World Neurosurg* **2023**. https://doi.org/10.1016/j.wneu.2023.08.042.
31. Kleebayoon, A.; Mungmunpuntipanitp, R.; Wiwanitkit, V. Chat GPT in Stereotactic Radiosurgery: Correspondence. *J Neurooncol* **2023**, *163* (3), 727–728. https://doi.org/10.1007/s11060-023-04375-7.
32. Ismail, A. M. A. Chat GPT in Tailoring Individualized Lifestyle-Modification Programs in Metabolic Syndrome: Potentials and Difficulties? *Ann Biomed Eng* **2023**. https://doi.org/10.1007/s10439-023-03279-x.
33. Arslan, S. Exploring the Potential of Chat GPT in Personalized Obesity Treatment. *Ann Biomed Eng* **2023**, *51* (9), 1887–1888. https://doi.org/10.1007/s10439-023-03227-9.
34. Zhou, Z. Evaluation of ChatGPT's Capabilities in Medical Report Generation. *Cureus* **2023**. https://doi.org/10.7759/cureus.37589.
35. Grewal, H.; Dhillon, G.; Monga, V.; Sharma, P.; Buddhavarapu, V. S.; Sidhu, G.; Kashyap, R. Radiology Gets Chatty: The ChatGPT Saga Unfolds. *Cureus* **2023**. https://doi.org/10.7759/cureus.40135.
36. Iftikhar, S.; Naz, I.; Zahra, A.; Zaidi, S. zainab Y. Report Generation of Lungs Diseases From Chest X-Ray Using NLP. *International Journal of Innovations in Science and Technology* **2022**, *3* (5), 223–233. https://doi.org/10.33411/ijist/2021030518.
37. Száva-Kováts, E. Unfounded Attribution of the "Half-Life" Index-Number of Literature Obsolescence to Burton and Kebler: A Literature Science Study. *Journal of the American Society for Information Science and Technology* **2002**, *53* (13), 1098–1105. https://doi.org/10.1002/asi.10105.
38. Gorbea-Portal, S.; Atrián-Salazar, M. L. Medición de La Obsolescencia de La Información En Revistas de Salud Pública de México. *Gac Med Mex* **2018**, *154* (3). https://doi.org/10.24875/GMM.18003293.
39. Sallam, M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. *Healthcare* **2023**, *11* (6), 887. https://doi.org/10.3390/healthcare11060887.
40. Grech, V.; Cuschieri, S.; Eldawlatly, A. Artificial Intelligence in Medicine and Research – the Good, the Bad, and the Ugly. *Saudi J Anaesth* **2023**, *17* (3), 401. https://doi.org/10.4103/sja.sja_344_23.
41. Tay, T. H. C. Response to: "Medical Teacher's First ChatGPT's Referencing Hallucinations: Lessons for Editors, Reviewers, and Teachers". *Med Teach* **2023**, 1. https://doi.org/10.1080/0142159X.2023.2245129.
42. Emsley, R. ChatGPT: These Are Not Hallucinations – They're Fabrications and Falsifications. *Schizophrenia* **2023**, *9* (1), 52. https://doi.org/10.1038/s41537-023-00379-4.
43. Masters, K. Medical Teacher ′s First ChatGPT's Referencing Hallucinations: Lessons for Editors, Reviewers, and Teachers. *Med Teach* **2023**, *45* (7), 673–675. https://doi.org/10.1080/0142159X.2023.2208731.
44. Frosolini, A.; Gennaro, P.; Cascino, F.; Gabriele, G. In Reference to "Role of Chat GPT in Public Health", to Highlight the AI's Incorrect Reference Generation. *Ann Biomed Eng* **2023**. https://doi.org/10.1007/s10439-023-03248-4.
45. Rawashdeh, B.; Kim, J.; AlRyalat, S. A.; Prasad, R.; Cooper, M. ChatGPT and Artificial Intelligence in Transplantation Research: Is It Always Correct? *Cureus* **2023**. https://doi.org/10.7759/cureus.42150.
46. Harrington, L. ChatGPT Is Trending: Trust but Verify. *AACN Adv Crit Care* **2023**, e1–e7. https://doi.org/10.4037/aacnacc2023129.
47. Copeland-Halperin, L. R.; O'Brien, L.; Copeland, M. Evaluation of Artificial Intelligence–Generated Responses to Common Plastic Surgery Questions. *Plast Reconstr Surg Glob Open* **2023**, *11* (8), e5226. https://doi.org/10.1097/GOX.0000000000005226.
48. Kim, Y.-W.; Barach, P.; Melzer, A. The Seoul Declaration: A Manifesto for Ethical Medical Technology. *Minimally Invasive Therapy & Allied Technologies* **2019**, *28* (2), 69–72. https://doi.org/10.1080/13645706.2019.1596956.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.