scientific reports

Check for updates

OPEN

Harnessing ChatGPT and GPT-4 for evaluating the rheumatology questions of the Spanish access exam to specialized medical training

Alfredo Madrid-García[1]✉, Zulema Rosales-Rosado[1,6], Dalifer Freites-Nuñez[1,6], Inés Pérez-Sancristóbal[1,6], Esperanza Pato-Cour[1,6], Chamaida Plasencia-Rodríguez[2,6], Luis Cabeza-Osorio[3,4,6], Lydia Abasolo-Alcázar[1], Leticia León-Mateos[1], Benjamín Fernández-Gutiérrez[1,5] & Luis Rodríguez-Rodríguez[1]

The emergence of large language models (LLM) with remarkable performance such as ChatGPT and GPT-4, has led to an unprecedented uptake in the population. One of their most promising and studied applications concerns education due to their ability to understand and generate human-like text, creating a multitude of opportunities for enhancing educational practices and outcomes. The objective of this study is twofold: to assess the accuracy of ChatGPT/GPT-4 in answering rheumatology questions from the access exam to specialized medical training in Spain (MIR), and to evaluate the medical reasoning followed by these LLM to answer those questions. A dataset, RheumaMIR, of 145 rheumatology-related questions, extracted from the exams held between 2010 and 2023, was created for that purpose, used as a prompt for the LLM, and was publicly distributed. Six rheumatologists with clinical and teaching experience evaluated the clinical reasoning of the chatbots using a 5-point Likert scale and their degree of agreement was analyzed. The association between variables that could influence the models' accuracy (i.e., year of the exam question, disease addressed, type of question and genre) was studied. ChatGPT demonstrated a high level of performance in both accuracy, 66.43%, and clinical reasoning, median (Q1–Q3), 4.5 (2.33–4.67). However, GPT-4 showed better performance with an accuracy score of 93.71% and a median clinical reasoning value of 4.67 (4.5–4.83). These findings suggest that LLM may serve as valuable tools in rheumatology education, aiding in exam preparation and supplementing traditional teaching methods.

The emergence of large language models (LLM) with remarkable performance such as ChatGPT, has led to an unprecedented uptake in the population, being the fastest-growing application in history[1], and with the potential to transform various domains, including medicine[2]. A clear example of this burgeoning interest is the notable surge in scientific research focusing on ChatGPT's role within the medical field. From January 1st, 2023 until July 10th, 2023, there have been over 768 publications indexed in PubMed, that contain the string "ChatGPT" and that delve into various aspects of this Generative Pre-trained Transformer (GPT) chatbot. Some of the topics of such publications include the potential impact of ChatGPT on clinical and translational medicine[3], clinical pharmacology[4], scientific communication[5] and medical writing[6], medical evidence summarization[7], patient

[1]Grupo de Patología Musculoesquelética, Hospital Clínico San Carlos, Instituto de Investigación Sanitaria del Hospital Clínico San Carlos (IdISSC), Prof. Martin Lagos S/N, 28040 Madrid, Spain. [2]Reumatología, Hospital Universitario La Paz-IdiPaz, Paseo de La Castellana, 261, 28046 Madrid, Spain. [3]Medicina Interna, Hospital Universitario del Henares, Avenida de Marie Curie, 0, 28822 Madrid, Spain. [4]Facultad de Medicina, Universidad Francisco de Vitoria, Carretera Pozuelo, Km 1800, 28223 Madrid, Spain. [5]Facultad de Medicina, Universidad Complutense de Madrid, Madrid, Spain. [6]These authors contributed equally: Zulema Rosales-Rosado, Dalifer Freites-Nuñez, Inés Pérez-Sancristóbal, Esperanza Pato-Cour, Chamaida Plasencia-Rodríguez and Luis Cabeza-Osorio. ✉email: alfredo.madrid@salud.madrid.org

question-answering[8], public health[9], ethical and legal[10], health policy making[11], plastic and colorectal surgery[12,13], doctor-patient communication[9], or drug-drug interaction and explainability[14].

In rheumatology, the use of artificial intelligence (AI) and LLM, such chatbots, is gaining relevance[15]. A query conducted on July 10th in PubMed, "Rheumatology AND ChatGPT" produced 12 results. For instance, in Ref.[16], the authors asked ChatGPT to draft an editorial about AI potentially replacing rheumatologists in editorial writing and discussed the ethical aspects and the future role of rheumatologists. In Ref.[17], the authors discussed the role of ChatGPT as an author, and the inherent terms associated with the concept of authorship such as accountability, bias, accuracy, and responsibility. Eventually, they considered refusing to recognize ChatGPT as an author, in accordance also to *Nature* announcement[18]. For their part, the authors in Ref.[19] provided an overview of the potential use of ChatGPT as a rheumatologist, its capabilities, and risks. They concluded that although chatbots will not be able to make clinical decisions for now; they will be able, for example, to design study protocols, streamline information accessibility and contribute to time-saving. A similar correspondence article highlighted five areas in which ChatGPT has the potential to assist in rheumatology care delivery, to note: patient drug education, medical imaging reporting, medical note-keeping for outpatient consultations, medical education and training and clinical audit and research. Authors in Ref.[20] used ChatGPT as an assistant tool to aid in writing a case report of a systemic lupus erythematosus patient. Recently, a study[21] aimed to assess the diagnostic accuracy of GPT-4 in comparison with rheumatologists. For that purpose, this LLM was instructed to name the top five differential diagnoses. Authors concluded that GPT-4 showed a higher accuracy for the top three overall diagnoses compared to the rheumatologist's assessment.

One of the most promising and studied applications for tools like ChatGPT lies within the realm of education. The ability of these language models to understand and generate human-like text creates a multitude of opportunities for enhancing educational practices and outcomes[22–26]. Within this field, one of the most extensively studied applications of ChatGPT is its ability and reasoning in answering medical examination questions. One of the most prominent studies was[27]. This study demonstrated how ChatGPT performed at or near the passing threshold for all three tests of the United States Medical Licensing Exam despite not having undergone any specialized training or reinforcement. However, many similar studies have emerged in recent months.

Given the range of studies that have subjected ChatGPT to different medical exam questions (e.g., multiple-choice, open-ended), the disparity in the results obtained, and the proficiency of ChatGPT in understanding context and generating detailed responses, we hypothesize that these tools could aid in the learning process of medical students, particularly in the study of rheumatology and musculoskeletal diseases.

Therefore, the objective of this study is twofold:

First, we seek to ascertain the accuracy of ChatGPT/GPT-4 in answering rheumatology questions from the access exam to specialized medical training in Spain, Médico Interno Residente (MIR).
Secondly, we aim to evaluate the clinical reasoning followed by ChatGPT/GPT-4 in answering those multiple-choice questions.

Methods

ChatGPT and GPT-4 as large language models

ChatGPT, an iteration of the Generative Pre-trained Transformer (GPT) model, is an AI chatbot, which belongs to the category of large language models (LLM), that was developed by OpenAI. These models are language models (i.e., a probabilistic distribution of word sequences) that rely on neural networks with millions of parameters. ChatGPT is trained on diverse and extensive data sources (e.g., books, websites, and so on) and exhibits a groundbreaking ability to generate relevant and context-aware responses, making it a promising tool in the medical education area. The training data cut-off was September 2021. This means that data after this date were not used for training. The model is defined by OpenAI as "Our fastest model, great for most everyday tasks".

On the other hand, GPT-4, is a large multimodal (i.e., accepts text and image as input) model, also developed by OpenAI, faster, with more parameters and better performance than ChatGPT. More details and a comparison between both LLM used in this work are available in Ref.[28]. The model is defined by OpenAI as "Our most capable model, great for tasks that require creativity and advanced reasoning". For the purpose of this study, we utilized the version of ChatGPT/GPT-4 that was released on May 3rd 2023, as documented in OpenAI's release notes[29]. This version was used to answer all the rheumatology questions from the MIR exams in our study.

MIR exam

The MIR is an annual competitive examination required for entry into specialist medical training in Spain. It is comprised of 210 questions from more than 30 competencies (i.e., pediatrics, hematology, ophthalmology), see the Supplementary Material Excel File "MIR Competencies", and follows a multiple-choice format (i.e., since the 2015–2016 academic year, the number of choices decreased from five to four), with only one correct answer. Each question on the exam typically presents a clinical case or a factual query, and the exam also includes image-based questions. The number of questions per competency varies per year, and both the exam and the answers are officially published by the Spanish Ministry of Health[30]. The questions from this examination serve as an invaluable resource for this study, as they represent standardized, expertly crafted questions designed to evaluate a comprehensive understanding of medical subjects, and, therefore, are leveraged to evaluate the accuracy and clinical reasoning of ChatGPT and GPT-4 when exposing them to the rheumatology discipline.

Inclusion criteria

Rheumatology-related questions from MIR exams published from 2009–2010 to 2022–2023 were included in this study. Questions assigned to other specialities, (e.g., pediatrics, orthopedics), that intersected with the field of rheumatology were also included.

On the other hand, questions containing images were excluded from the analysis. This decision was taken because of the current limitations of ChatGPT, as it is primarily a text-based model and does not possess the capability to process or interpret image data. GPT-4 allows images as input data, but this feature was not publicly available at the time of this research. Hence, any questions in the MIR exams that were dependent on visual information, such as graphs, pictures, or clinical image data (e.g., X-ray), were not included. Finally, questions invalidated by the Spanish Ministry of Health were also excluded from the analysis.

Methodology

Questions from the MIR exams were used as prompts. For each question provided to ChatGPT/GPT-4, the sentence "Justify your answer" was added at the end. The responses generated by ChatGPT and GPT-4 to the rheumatology questions were evaluated by six independent rheumatologists. Three of them, ZRR, LCO, and CPR, in addition to being practising clinicians, are MIR training professors. The length of the text generated by the models was not artificially limited.

The medical experts evaluated the clinical reasoning of the chatbots followed in each of the responses. Their evaluation was based on a 1–5 scale, where a score of 5 indicates that the reasoning was entirely correct and flawless, while a score of 1 signifies that the reasoning was inconsistent or contained significant errors. Intermediate scores were used to denote minor errors in reasoning and the severity of these errors was reflected in the score; see Table 1. The final clinical reasoning score was assigned following a majority vote approach. In case of a tie, the worst score was chosen. The evaluators were also asked to justify the score given to ChatGPT/GPT-4 clinical reasoning when necessary. After that, each question was categorized based on the type of disease being asked about, and classified into factual or clinical case questions.

For this study, we solely relied on the initial responses generated by ChatGPT/GPT-4, without employing the "regenerate response" function. The questions were prompted in Spanish, exactly as they were extracted from the exam. Nonetheless, an English translation of both the question and the clinical reasoning by the LLM was obtained with *DeepL*. In cases where the answers provided by the models were not singular, a new prompt was used to ask for a single and unique response with the following instruction: "If you had to choose one of the answers, the most relevant one, which would it be?".

The clinical reasoning of the models was evaluated by the medical experts in Spanish.

A questionnaire was provided to each evaluator to assess the overall performance of the language models and their suitability for use in education and clinical practice. This questionnaire comprises seven free-text questions and can be found in the Supplementary Material "Questionnaire".

Finally, the exam questions and chatbot answers lacked identifiable patient information.

Variables

Two main variables were considered for evaluation:

- Accuracy, defined as the match between the official MIR question response and the chosen option by ChatGPT/GPT-4.
- Score assigned to the clinical reasoning of ChatGPT/GPT-4 by the six evaluators.

Covariables to be considered in this study included: year of the exam question, type of question, patient's gender, pathology that the question addresses, and chatbot model used.

Statistical analysis

Dichotomous and categorical variables were summarized using proportions. Continuous variables were summarized using the median and the first and third quartiles (Q1–Q3). The distribution of correct answers (i.e., accuracy) among the different covariates was analyzed using chi-square or Fisher's test, depending on the number of events. Differences between LLM, in terms of accuracy, were evaluated using McNemar's test.

The degree of agreement in the score assigned to the clinical reasoning by the evaluators was analyzed using Krippendorff's alpha coefficient, Kendall's coefficient with and without tie correction coefficients[31] and Gwet's AC2 coefficient[32]. The Kappa-Fleiss coefficient was not used as raters are considered unique (i.e., all evaluators

Score	Reasoning
1	Wrong answer and completely incorrect reasoning
2	Wrong answer and not completely incorrect reasoning (e.g., some information can be valuable)
3	Wrong answer and acceptable reasoning OR correct answer but poor reasoning
4	Correct answer and acceptable reasoning, but with important details not commented on or mentioned
5	Correct answer and correct and complete reasoning, without important details uncommented or unmentioned

Table 1. Template of the reasoning score given to the evaluators.

justify the reasoning of all questions). The final clinical reasoning score given to each question was determined by a majority vote among evaluators. In the event of a tie, the worst score was chosen.

Differences between LLM, in terms of clinical reasoning, were evaluated using Wilcoxon signed-rank test. The effect of covariates on the clinical reasoning score was studied using ordinal logistic regression.

R version 4.3.1 was used to perform the statistical analysis.

Ethics board approval

As suggested in Ref.[10], the Hospital Clínico San Carlos (HCSC) Ethics Review Board evaluated this project, 23/370-E, and stated that this committee was not competent to evaluate studies of this type, since they do not encompass human subjects, or the use of biological samples, or personal data.

Terms of use of the services used in this work

The use of language models in this study complies with the established terms of service of OpenAI (https://openai.com/policies/terms-of-use), Google BARD (https://support.google.com/bard/answer/13594961?hl=en) and Antrophic (https://console.anthropic.com/legal/terms).

Preprint

A previous draft of this work was published as a preprint and can be found at: Madrid-García, A. *et al*. Harnessing ChatGPT and GPT-4 for evaluating the rheumatology questions of the Spanish access exam to specialized medical training. Preprint at medRxiv, 2023-07 (2023).

Results

The questions evaluated by ChatGPT/GPT-4, the answer by both systems, as well as the official answer, the medical experts' evaluation of the clinical reasoning, the year, genre, type of question, whether the question was invalidated or not, the type of disease being asked about, the English translation of the questions and the clinical reasoning are shown in the RheumaMIR dataset accessible through *Zenodo*[33].

Description

After applying the inclusion criteria, 143 questions from 14 MIR exams remained (i.e., academic years 2009–2010 to 2022–2023). Table 2 shows the dataset characteristics. The median number of questions (Q1-Q3) per year is 11 (9.25–12) and the most prevalent disease being asked about is vasculitis. Most of the questions, 65.73%, were clinical cases, and the sex of the clinical case subjects was evenly distributed.

Accuracy

Out of 143 questions, GPT-4 correctly answered 134 (93.71%), demonstrating a high level of accuracy. On the other hand, ChatGPT accurately answered 95 questions (66.43%), indicating a somewhat lower level of performance in comparison to GPT-4 (i.e., McNemar's test p-value = 1.17×10^{-09}).

ChatGPT did not correctly answer any of the questions that GPT-4 failed to answer correctly. Moreover, out of the nine questions that GPT-4 got wrong, seven of them also matched the answer given by ChatGPT. Table 3 shows the number and percentage of errors per model and covariate. Eventually, none of the covariates were associated with the number of errors, neither in GPT-4 nor in ChatGPT.

Clinical reasoning

The Krippendorff's alpha coefficient, Kendall's coefficients with and without tie correction and Gwet's AC2 coefficient, for the clinical reasoning of GPT-4 considering the six evaluators, were 0.225, 0.452, 0.269 and 0.924, and for ChatGPT, 0.624, 0.783, 0.636 and 0.759. A more detailed analysis of the main differences between evaluators' scores and an error analysis discussing the questions failed by GPT-4 can be found in the Supplementary Material File 'Evaluator agreement' and 'Error analysis' sections and Supplementary Material Excel File Agreement Coefficients.

The median (Q1–Q3) value of the average score for the clinical reasoning of GPT-4 was 4.67 (4.5–4.83), while for ChatGPT was 4.5 (2.33–4.67). There exist statistically significant differences in the clinical reasoning score, after applying the majority vote, of both LLM (i.e., Wilcoxon signed rank test p-value = 4.47×10^{-09}).

Regarding the covariates, there were no statistically significant differences in the clinical reasoning score for GPT-4/ChatGPT after applying ordinal logistic regression models. Figure 1 shows the proportion of scores given by the evaluators, grouped by score. Except for one evaluator, Evaluator 6, the most repeated score given was 5 for both models, with a small percentage of low scores (i.e., 1, 2, 3). The comparison of the scores given by the evaluators to the clinical reasoning of ChatGPT and GPT-4, grouped by evaluation, is shown in Supplementary Fig. 1. For both figures, the majority vote statistics are shown.

On its behalf, Fig. 2 shows the proportion between the clinical reasoning score and the disease addressed in the questions. Supplementary Figs. 2–4, show the proportion between the clinical reasoning score and the year, the type of question, and the genre. There does not appear to be a clear trend between the reasoning score and the covariates shown in these plots. Finally, the completed questionnaires can be found in the Supplementary Material 'Questionnaire'. The evaluators concur on:

- The potential usefulness of this tool, particularly in creating educational content, albeit under expert supervision.

Variable	n (%)
Year/number of questions	
2009–2010	12 (8.39%)
2010–2011	11 (7.69%)
2011–2012	11 (7.69%)
2012–2013	12 (8.39%)
2013–2014	5 (3.50%)
2014–2015	11 (7.69%)
2015–2016	12 (8.39%)
2016–2017	9 (6.29%)
2017–2018	10 (6.99%)
2018–2019	6 (4.20%)
2019–2020	11 (7.69%)
2020–2021	15 (10.49%)
2021–2022	6 (4.20%)
2022–2023	12 (8.39%)
Disease	
Scleroderma	8 (5.59%)
Microcrystalline arthritis	9 (6.29%)
Infective arthritis	11 (7.69%)
Rheumatoid arthritis	12 (8.39%)
Spondyloarthropathies	14 (9.79%)
Systemic lupus erythematosus	16 (11.19%)
Bone metabolism	20 (13.99%)
Vasculitis	23 (16.08%)
Others[‡]	30 (20.98%)
Type of question	
Factual question	49 (34.27%)
Clinical case	94 (65.73%)
Sex	
No sex/newborn	2 (1.40%)
Female	42 (29.37%)
Not applicable	49 (34.27%)
Male	50 (34.97%)

Table 2. Dataset description. [‡]Including amyloidosis and autoinflammatory syndromes, fibromyalgia, IgG4-related disease inflammatory myopathy, osteoarthritis, others, sarcoidosis, Sjögren's syndrome.

- The language used in the responses may lack technical precision and could be suitable for students, but not in other scenarios (i.e., official medical documentation).
- ChatGPT/GPT-4 models are unaware of the limitations and scope of their knowledge, sometimes justifying facts in a tortuous manner and potentially misleading the reader.

Discussion

In this study, we have evaluated the accuracy and clinical reasoning of two LLM in answering rheumatology questions from Spanish official medical exams. To our knowledge, this is the first study to evaluate the usefulness of LLM applied to the training of medical students with a special focus on rheumatology.

The ability of GPT-4 to answer questions with high accuracy and sound clinical reasoning is remarkable. This could make such models valuable learning tools for medical students. However, ChatGPT/GPT-4 LLM are only the first models that have reached the public in the rapidly expanding field of LLM chatbots. At present, a myriad of additional models are under development. Some of these nascent models are not only pre-trained in biomedical texts[34,35], but are also specifically designed for a broad range of tasks (e.g., text summarization, question-answering and so on).

Studies with a similar objective to this one have been conducted. For example, a Spanish study[36], evaluated ChatGPT's ability to answer questions from the 2022 MIR exam. In this cross-sectional and descriptive analysis, 210 questions from the exam were entered into the model. ChatGPT correctly answered 51.4% of the questions. This resulted in a 7688 position, slightly below the median of the population tested but above the passing score.

In another research[37], the proficiency of ChatGPT in answering higher-order thinking questions related to medical biochemistry, including 11 competencies such as basic biochemistry, enzymes, chemistry and metabolism of carbohydrates, lipids and proteins, oncogenesis, and immunity, was studied. Two-hundred questions

Variable	ChatGPT error	GPT-4 error
Year		
2009–2010	3 (25.00%)	1 (8.33%)
2010–2011	4 (36.36%)	2 (18.18%)
2011–2012	4 (36.36%)	1 (9.09%)
2012–2013	3 (25.00%)	–
2013–2014	2 (40.00%)	–
2014–2015	3 (27.27%)	–
2015–2016	4 (33.33%)	–
2016–2017	3 (33.33%)	–
2017–2018	3 (30.00%)	–
2018–2019	4 (66.67%)	2 (33.33%)
2019–2020	4 (36.36%)	1 (9.09%)
2020–2021	6 (40.00%)	2 (13.33%)
2021–2022	1 (16.67%)	–
2022–2023	4 (33.33%)	–
p-value	0.979	0.233
Disease		
Scleroderma	2 (25.00%)	1 (12.50%)
Microcrystalline arthritis	6 (66.67%)	–
Infective arthritis	4 (36.36%)	–
Rheumatoid arthritis	3 (25.00%)	1 (8.33%)
Spondyloarthropathies	2 (14.29%)	1 (7.14%)
Systemic lupus erythematosus	6 (37.50%)	–
Bone metabolism	5 (25.00%)	1 (5.00%)
Vasculitis	10 (43.48%)	3 (13.04%)
Others	10 (33.33%)	2 (6.67%)
p-value	0.355	0.816
Type of question		
Factual	18 (36.73%)	6 (12.24%)
Clinical case	30 (31.91%)	3 (3.19%)
p-value	0.580	0.063
Sex		
Women	14 (33.33%)	1 (2.38%)
Men	15 (30.00%)	2 (4.00%)
p-value	1	0.823

Table 3. Number of errors per model and covariate.

were randomly chosen from an institution's question bank and classified according to the Competency-Based Medical Education. The answers were evaluated by two expert biochemistry academicians on a scale of zero to five. ChatGPT obtained a median score of 4 out of 5, with oncogenesis and immunity competition having the lowest score and basic biochemistry the competition with the highest.

Research of a similar nature was conducted in Ref.[38]. In this study, the authors appraised the capability of ChatGPT in answering first- and second-order questions on microbiology (e.g., general microbiology and immunity, musculoskeletal system, skin and soft tissue infections, respiratory tract infections and so on) from the Competency Based Medical Education curriculum. A total of 96 essay questions were reviewed for content validity by an expert microbiologist. Subsequently, ChatGPT responses were evaluated on a scale of 1 to 5, with five being the highest score, by three microbiologists. A median score of 4.04 was achieved.

On the other hand, ChatGPT was tested on the Plastic Surgery In-Service examinations from 2018 to 2022 and its performance was compared to the national average performance of plastic surgery residents[39]. Out of 1129 questions, ChatGPT answered 630 (55.8%) correctly. When compared with the performance of plastic surgery residents in 2022, ChatGPT ranked in the 49th percentile for first-year residents, but its performance fell significantly among residents in higher years of training, dropping to the 0th percentile for 5th and 6th-year residents.

Another study was conducted by researchers in Ref.[40], who aimed to assess whether ChatGPT could score equivalently to human candidates in a virtual objective structured clinical examination in obstetrics and gynecology. Seven structured examination questions were selected, and the responses of ChatGPT were compared to the responses of two human candidates and evaluated by fourteen qualified examiners. ChatGPT received an average score of 77.2%, while the average historical human score was 73.7%.

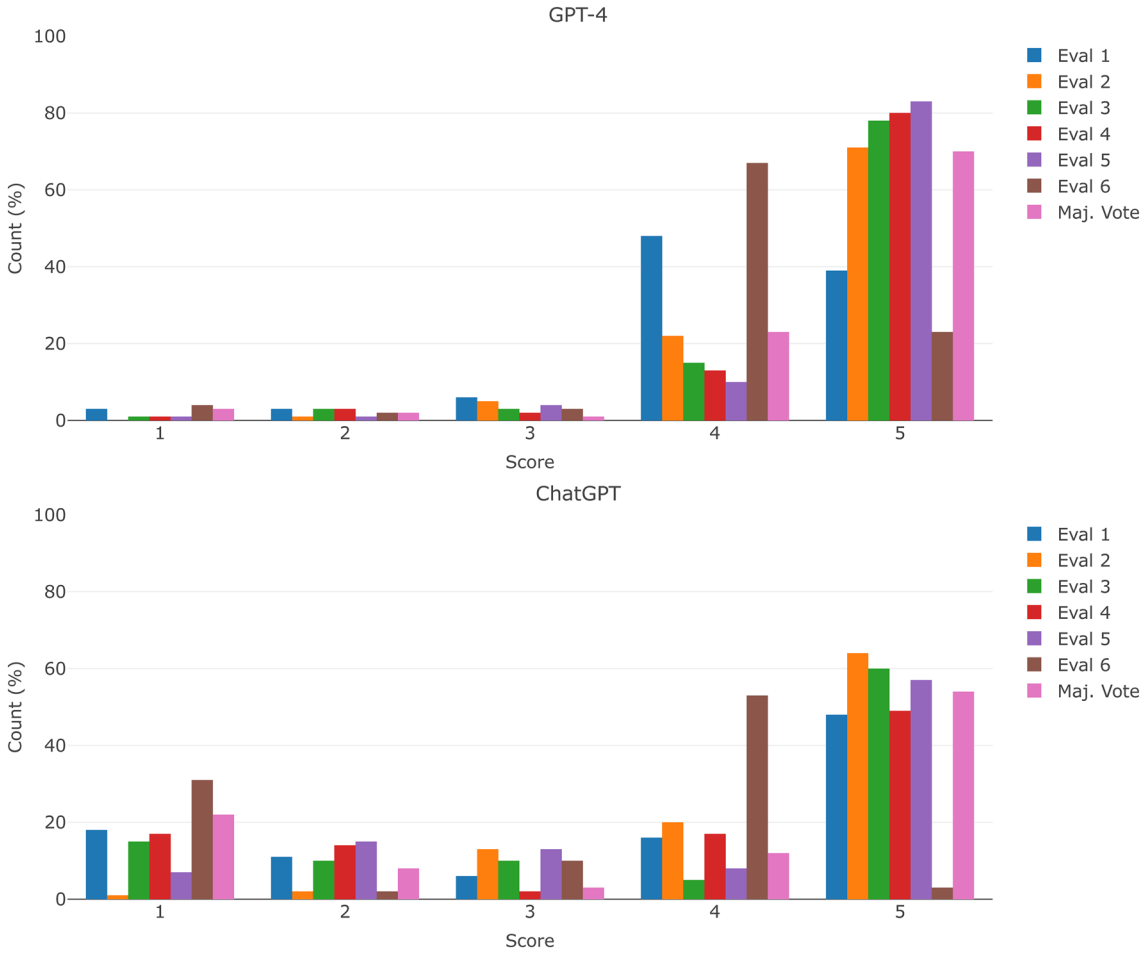

Figure 1. Distribution of the scores given by the evaluators.

Moreover, the authors in Ref.[41] instructed ChatGPT to deliver concise answers to the 24-item diabetes knowledge questionnaire, consisting of a clear "Yes" or "No" response, followed by a concise rationale comprising two sentences for each question. The authors found that ChatGPT successfully answered all the questions.

In Ref.[42], the researchers were interested in evaluating the performance of ChatGPT on open-ended clinical reasoning questions. Therefore, fourteen multi-part cases were selected from clinical reasoning exams administered to first and second-year medical students and provided to ChatGPT. Each case was comprised of 2–7 open-ended questions and was shown to ChatGPT twice. ChatGPT achieved or surpassed the pre-established passing score of 70% in 43% of the runs (12 out of 28), registering an average score of 69%.

Some studies showed remarkable performance, for instance, a research study evaluated the performance of ChatGPT in medical physiology university examination of phase I MBBS[43]. In this investigation, ChatGPT correctly answered 17 out of 20 multiple-choice questions, while providing a comprehensive explanation for each one. On their side, researchers in Ref.[44] proposed a four-grading system to classify the answers of ChatGPT, to note, comprehensive, correct but inadequate, mixed with correct and incorrect/outdated data, and completely incorrect. ChatGPT showed a 79% and a 74% of accuracy when answering questions related to cirrhosis and hepatocellular carcinoma. However, only the 47% and 41% of the answers were classified as comprehensive.

Conversely, in another research[45] in which ChatGPT was exposed to the family medicine course's multiple-choice exam of Antwerp University, only 2/125 students performed worse than ChatGPT. Since the questions were prompted in Dutch language, the potential correlation between ChatGPT's low performance and the proportion of Dutch texts used in its training could be a factor worth considering for this discordant result.

Another study, Ref.[46], evaluated ChatGPT's performance on standardized admission tests in the United Kingdom, including the BioMedical Admissions Test (BMAT), Test of Mathematics for University Admission (TMUA), Law National Aptitude Test (LNAT), and Thinking Skills Assessment (TSA). A dataset of 509 multiple-choice questions from these exams, ranging from 2019 to 2022 was used. The results varied among specialities. For BMAT, the percentage of correct answers varied from 5 to 66%, for TMUA varied from 11 to 22%, for LNAT from 36 to 53%; and for TSA from 42 to 60%. The authors concluded that while ChatGPT demonstrated potential as a supplemental tool for areas assessing aptitude, problem-solving, critical thinking, and reading comprehension, it showed limitations in scientific and mathematical knowledge and applications.

The results shown by most of these studies are in line with our results, the average score of ChatGPT is between 4 and 5 (on a scale of five elements) when answering medical-related questions. However, in these

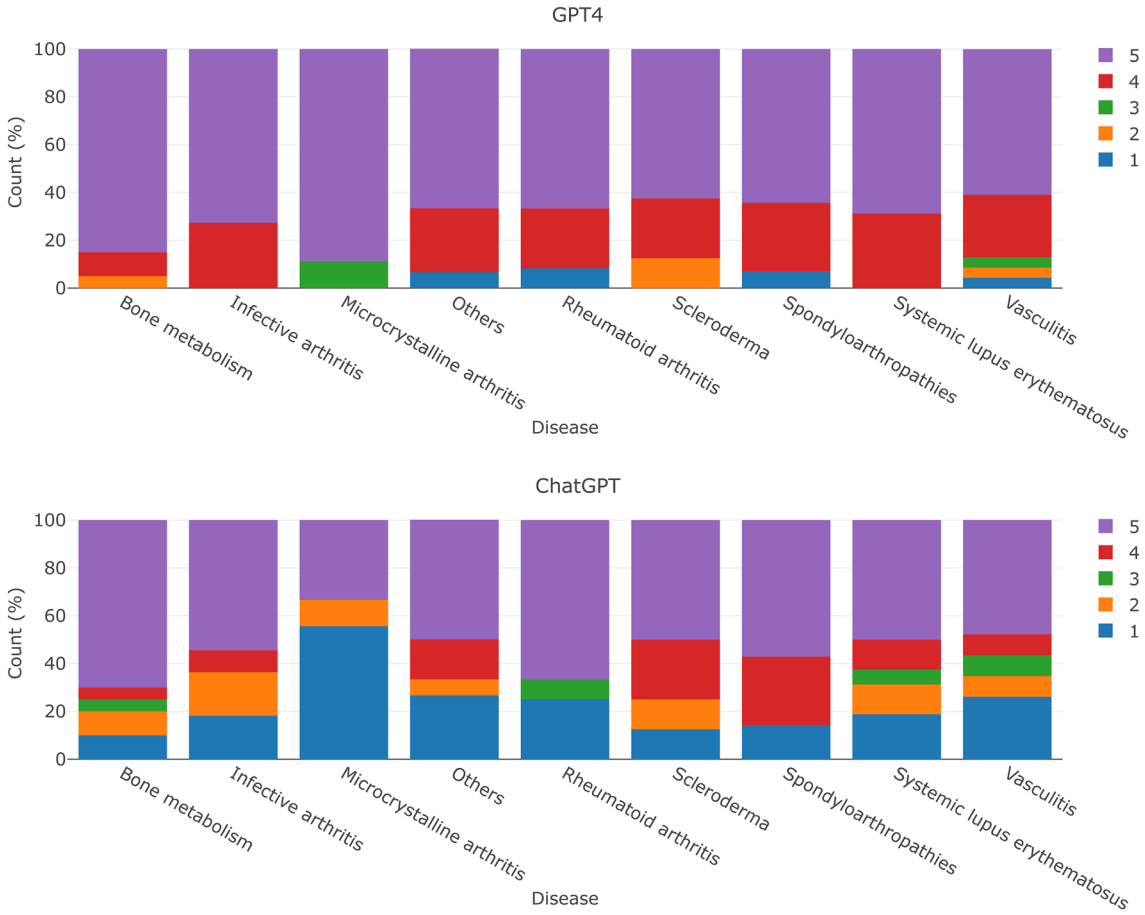

Figure 2. Clinical reasoning score according to the disease addressed in the question after taking the majority vote. In case of a tie, the worst score was chosen.

studies, GPT-4 performance was not evaluated. Based on our results, we can postulate that there would be an increase in accuracy in comparison to those obtained by ChatGPT. In addition, to solve some of the limitations identified by our evaluators, such as the employment of a language that can lack technical precision by the models, LLM could perform better if trained or fine-tuned with biomedical texts.

A large part of the concerns and doubts that arise from using these models are due to regulatory and ethical issues. Some of the ethical dilemmas have been highlighted in Ref.[47]. For instance, the authors pointed out that LLM reflect any false, outdated, and biased data from which the model was trained, and that they could not reflect the latest guidelines. Some authors have pointed out the risks of perpetuating biases if the model has been trained on biased data[48,49] Other authors go further and declare that these types of models should not be used in clinical practice[50]. The motivation behind this statement lies in the presence of biases such as clinician bias, which may exacerbate racial-ethnic disparities, or "hallucinations" meaning that ChatGPT produces high levels of confidence in its output even when insufficient or masked information in the prompt. According to the authors, this phenomenon could lead users to place unwavering trust in the output of chatbots, even if it contains unreliable information. This was also pointed out by our evaluators, as shown in the Supplementary Material 'Questionnaire'. The firmness with which these models justify erroneous reasoning may limit their potential usefulness. Eventually, the authors also claimed that these models are susceptible to "Falsehood Mimicry", that is, the model will attempt to generate an output that aligns with the user's assumption rather than clarifying questions. Falsehood mimicry and hallucinations may limit the potential use of these models as diagnostic decision support systems (DDSS). For instance, a clinical trial[51] compared the diagnostic accuracy of medical students regarding the typical rheumatic diseases, with and without the use of a DDSS and concluded that no significant advantage was observed from the use of the DDSS. Moreover, researchers reported that students accepted false DDSS diagnostic suggestions in a substantial number of situations. This phenomenon could be exacerbated when using LLM, and therefore should be study with caution.

In our study, due to the nature of the questions, we were unable to assess racial or ethnic disparities. However, we did not find any gender bias when considering the clinical case questions. Finally, a relevant study that looks in depth at the biases that can arise when using LLMs can be found in Ref.[52].

Regarding regulation, the arrival of these models has led to greater efforts being made to regulate the use of AI. The EU AI Act[53] is a good example of this. According to our results, in these early stages of LLM, the corpus used for training, as well as the content generated by them, should be carefully analyzed.

During the writing of this manuscript, new articles have emerged. For instance, in Ref.[54] authors studied the performance of ChatGPT when introducing the most searched keywords related to seven rheumatic and musculoskeletal diseases. The content of each answer was evaluated in terms of usefulness for patients with the ChatGPT in a scale from 1 to 7 by two raters.

Finally, further analysis is needed to explore these observations and understand their implications for the development and use of AI in medical education and practice.

Limitations

- Two chatbots were primarily used in this study, ChatGPT and GPT-4, both owned by OpenAI. However, other LLM such as BARD or Med-PaLM2 by Google, Claude 2 by Anthropic or LLaMA and LIMA by Meta are in development. Some of them are publicly available. To provide a better overview of other LLM, the accuracy of BARD (60.84%) and Claude 2 (79.72%) was calculated and compared against ChatGPT/GPT-4 in the Supplementary Material File Section 'LLM comparison' and Supplementary Fig. 5.
- The Krippendorff's alpha coefficient and Kendall's coefficient of agreement oscillates between 0.225 and 0.452 for the clinical reasoning of GPT-4, although the most repeated score of five out of six evaluators is 5 (the percentage of four and five scores oscillates between 87.41% and 93.70%), see Fig. 1. This phenomenon is known as "Kappa paradoxes" (i.e., 'high agreement, but low reliability')[55,56], and tends to appear in skewed distributions such as the one presented in this study. More details can be found in Ref.[31]. In this study, since the ChatGPT clinical reasoning score distribution is less skewed, the reliability coefficient values are higher, between 0.624 and 0.783, than the ones obtained with GPT-4. However, when considering the Gwet's AC2 coefficient, the trend is reversed, 0.924 vs. 0.759, with higher inter-rater agreement in GPT-4 compared to ChatGPT. These large differences between interrater reliability indices have been observed in simulation studies with skewed distributions[57].
- To ensure reproducibility and facilitate the comparison of results, each question could have been submitted twice to ChatGPT/GPT-4, a strategy supported by previous research endeavours[44], and a default feature of BARD. However, in the tests we conducted, the responses were consistent across iterations and this would have doubled the workload of evaluators, so we chose to include more questions from a single run, rather than fewer questions run multiple times. In addition, according to Ref.[44], the 90.48% of "regenerated questions" produced two similar responses with similar grading.
- The format of each question could have been transformed from multiple choice to open-ended. With this approach, it could have been possible to delve deeper into ChatGPT's clinical reasoning. As explained in the previous point, this would have doubled the workload. Additionally, there are no open questions in the MIR exams.
- When conducting such studies, it is crucial to consider the evolution of knowledge over time. Evaluating older questions with models trained on recent data may reveal disparities compared to previously accepted and conventional beliefs. Therefore, accounting for this temporal aspect is essential to ensure the accuracy and relevance of the study findings. Moreover, although not extensively explored in this research, one of the key concepts when using LLM chatbots is what is known as the prompt or input text that is entered into the model. Depending on how well-defined the prompt is, the results can vary significantly. In this study, we tried to adhere closely to the official question while minimizing any unnecessary additional text.
- Another identified limitation of the study is the absence of medical students evaluation for the models' clinical reasoning. This would have allowed us to determine whether students can recognize the main issues discussed above (e.g., bias, falsehood mimicry, hallucinations), and analyze to what extent these limitations may affect their usefulness.
- One of the main criticisms of the evaluators in assessing the LLM response was the use of non-technical language. This could have been remedied, in part, by using prompt engineering, this is, by modifying the initial input of the model and asking for a more technical response.
- We have explored the performance of GPT-4/ChatGPT in Spanish questions. However, different authors have suggested that such performance is language-dependent[58]. Hence, special caution should be taken when extrapolating the results to other languages.

Conclusion

The accuracy of ChatGPT and GPT-4 in answering the rheumatology questions of the Spanish access exam to specialized medical training is high, as well as the clinical reasoning. Nevertheless, discerning the veracity of such reasoning can pose a challenge for students in situations where the LLM experiences failures, given the comprehensive and seemingly accurate elaboration present in erroneous responses. Hence, extreme caution should be exercised when using these models as teaching aids and even greater diligence should be taken when using them in clinical practice. However, these kinds of models hold the potential to serve as a valuable asset in the development of pedagogical materials, subject to rigorous expert evaluation.

Data availability

The data that support the findings of this study are openly available in Zenodo at: Alfredo Madrid García, Zulema Rosales Rosado, Dalifer Freites Núñez, Inés Pérez San Cristobal, Esperanza Pato Cour, Chamaida Plasencia Rodríguez, & Luis Cabeza Osorio. (2023). RheumaMIR (3.0.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.10204293.

Received: 16 October 2023; Accepted: 8 December 2023
Published online: 13 December 2023

References

1. Dennean, K., Gantori, S., Limas, D. K., Pu, A. & Gilligan, R. *Let's Chat About ChatGPT*. https://www.ubs.com/global/en/wealth-management/our-approach/marketnews/article.1585717.html (2023).
2. Biswas, S. ChatGPT and the future of medical writing. *Radiology* **307**, 3312 (2023).
3. Xue, V. W., Lei, P. & Cho, W. C. The potential impact of ChatGPT in clinical and translational medicine. *Clin. Transl. Med.* **13**, 1216 (2023).
4. Krumborg, J. R. *et al.* ChatGPT: First glance from a perspective of clinical pharmacology. *Basic Clin. Pharmacol. Toxicol.* **133**, 3–5 (2023).
5. Huang, J. & Tan, M. The role of ChatGPT in scientific communication: Writing better scientific review articles. *Am. J. Cancer Res.* **13**, 1148–1154 (2023).
6. Biswas, S. Passing is great: Can ChatGPT conduct USMLE exams? *Ann. Biomed. Eng.* **51**, 1885–1886 (2023).
7. Tang, L. *et al.* Evaluating large language models on medical evidence summarization. *NPJ Digit. Med.* **6**, 158 (2023).
8. Lee, T.-C. *et al.* ChatGPT answers common patient questions about colonoscopy. *Gastroenterology* **165**, 509–511 (2023).
9. He, Y. *et al.* Can ChatGPT/GPT-4 assist surgeons in confronting patients with Mpox and handling future epidemics? *Int. J. Surg.* **109**, 2544–2548 (2023).
10. da Silva, J. A. T. Is institutional review board approval required for studies involving ChatGPT? *Am. J. Obstet. Gynecol. MFM* **5**, 101005 (2023).
11. Sifat, R. I. ChatGPT and the future of health policy analysis: Potential and pitfalls of using ChatGPT in policymaking. *Ann. Biomed. Eng.* **51**, 1357–1359 (2023).
12. Kang, Y., Xia, Z. & Zhu, L. When ChatGPT meets plastic surgeons. *Aesthetic Plast. Surg.* **47**, 2190–2193 (2023).
13. Li, W., Zhang, Y. & Chen, F. ChatGPT in colorectal surgery: A promising tool or a passing fad? *Ann. Biomed. Eng.* **51**, 1892–1897 (2023).
14. Juhi, A. *et al.* The capability of ChatGPT in predicting and explaining common drug–drug interactions. *Cureus*. https://doi.org/10.7759/cureus.36272 (2023).
15. Madrid-García, A. *et al.* Understanding the role and adoption of artificial intelligence techniques in rheumatology research: An in-depth review of the literature. *Semin. Arthritis Rheum.* **61**, 152213 (2023).
16. Verhoeven, F., Wendling, D. & Prati, C. ChatGPT: When artificial intelligence replaces the rheumatologist in medical writing. *Ann. Rheum. Dis.* **82**, 1015–1017 (2023).
17. Solomon, D. H. *et al.* Artificial intelligence, authorship, and medical publishing. *Arthritis Rheumatol.* **75**, 867–868 (2023).
18. Nature editorial. Tools such as ChatGPT threaten transparent science; here are our ground rules for their use. *Nature* **613**, 612 https://www.nature.com/articles/d41586-023-00191-1 (2023).
19. Hügle, T. The wide range of opportunities for large language models such as ChatGPT in rheumatology. *RMD Open* **9**, e003105 (2023).
20. Jansz, J., Manansala, M. J. & Sweiss, N. J. Treatment of periorbital edema in a patient with systemic lupus erythematosus during pregnancy: A case report written with the assistance of ChatGPT. *Cureus*. https://doi.org/10.7759/cureus.36302 (2023).
21. Krusche, M., Callhoff, J., Knitza, J. & Ruffer, N. Diagnostic accuracy of a large language model in rheumatology: Comparison of physician and ChatGPT-4. *Rheumatol. Int.* https://doi.org/10.1007/s00296-023-05464-6 (2023).
22. Grabb, D. ChatGPT in medical education: A paradigm shift or a dangerous tool? *Acad. Psychiatry* **47**, 439–440 (2023).
23. van de Ridder, J. M. M., Shoja, M. M. & Rajput, V. Finding the place of ChatGPT in medical education. *Acad. Med.* **98**, 867–867 (2023).
24. Munaf, U., Ul-Haque, I. & Arif, T. B. ChatGPT: A helpful tool for resident physicians? *Acad. Med.* **98**, 868–869 (2023).
25. Feng, S. & Shen, Y. ChatGPT and the future of medical education. *Acad. Med.* **98**, 867–868 (2023).
26. Seetharaman, R. Revolutionizing medical education: Can ChatGPT boost subjective learning and expression? *J. Med. Syst.* **47**, 61 (2023).
27. Kung, T. H. *et al.* Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS Digital Health* **2**, e0000198 (2023).
28. OpenAI. GPT-4. Preprint at (2023).
29. OpenAI. ChatGPT—Release notes. Preprint at (2023).
30. de España, M. D. & Sanidad, G. BOE-A-2022-14414. II. Autoridades y personal B. Oposiciones y concursos. Preprint at https://www.boe.es/boe/dias/2022/09/02/pdfs/BOE-A-2022-14414.pdf (2022).
31. Feng, G. C. Mistakes and how to avoid mistakes in using intercoder reliability indices. *Methodology* **11**, 13–22 (2015).
32. Gwet, K. L. Computing inter-rater reliability and its variance in the presence of high agreement. *Br. J. Math. Stat. Psychol.* **61**, 29–48 (2008).
33. García, A. M. RheumaMIR. Preprint at 10.5281/zenodo.8153291 (2023).
34. Jin, Q., Yang, Y., Chen, Q. & Lu, Z. GeneGPT: Augmenting large language models with domain tools for improved access to biomedical information. Preprint at (2023).
35. Wu, C., Zhang, X., Zhang, Y., Wang, Y. & Xie, W. PMC-LLaMA: Further finetuning LLaMA on medical papers. Preprint at (2023).
36. Carrasco, J. P. *et al.* ¿Es capaz "ChatGPT" de aprobar el examen MIR de 2022? Implicaciones de la inteligencia artificial en la educación médica en España. *Rev. Esp. Educ. Méd.* **4**, 1 (2023).
37. Ghosh, A. & Bir, A. Evaluating ChatGPT's ability to solve higher-order questions on the competency-based medical education curriculum in medical biochemistry. *Cureus*. https://doi.org/10.7759/cureus.37023 (2023).
38. Das, D. *et al.* Assessing the capability of ChatGPT in answering first- and second-order knowledge questions on microbiology as per competency-based medical education curriculum. *Cureus*. https://doi.org/10.7759/cureus.36034 (2023).
39. Humar, P., Asaad, M., Bengur, F. B. & Nguyen, V. ChatGPT is equivalent to first-year plastic surgery residents: Evaluation of ChatGPT on the plastic surgery in-service examination. *Aesthet. Surg. J.* **43**, 1085–1089 (2023).
40. Li, S. W. *et al.* ChatGPT outscored human candidates in a virtual objective structured clinical examination in obstetrics and gynecology. *Am. J. Obstet. Gynecol.* **229**, e1–e12 (2023).
41. Nakhleh, A., Spitzer, S. & Shehadeh, N. ChatGPT's response to the diabetes knowledge questionnaire: Implications for diabetes education. *Diabetes Technol. Ther.* **25**, 571–573 (2023).
42. Strong, E. *et al.* Performance of ChatGPT on free-response, clinical reasoning exams. *MedRxiv*. https://doi.org/10.1101/2023.03.24.23287731 (2023).
43. Subramani, M., Jaleel, I. & Krishna Mohan, S. Evaluating the performance of ChatGPT in medical physiology university examination of phase I MBBS. *Adv. Physiol. Educ.* **47**, 270–271 (2023).
44. Yeo, Y. H. *et al.* Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. *Clin. Mol. Hepatol.* **29**, 721–732 (2023).
45. Morreel, S., Mathysen, D. & Verhoeven, V. Aye, AI! ChatGPT passes multiple-choice family medicine exam. *Med. Teach.* **45**, 665–666 (2023).

46. Giannos, P. & Delardas, O. Performance of ChatGPT on UK standardized admission tests: Insights from the BMAT, TMUA, LNAT, and TSA examinations. *JMIR Med. Educ.* **9**, e47737 (2023).
47. Beltrami, E. J. & Grant-Kels, J. M. Consulting ChatGPT: Ethical dilemmas in language model artificial intelligence. *J. Am. Acad. Dermatol.* https://doi.org/10.1016/j.jaad.2023.02.052 (2023).
48. Wang, C. *et al.* Ethical considerations of using ChatGPT in health care. *J. Med. Internet Res.* **25**, e48009 (2023).
49. Ferrara, E. Should ChatGPT be biased? Challenges and risks of bias in large language models. *First Monday.* https://doi.org/10.5210/fm.v28i11.13346 (2023).
50. Au Yeung, J. *et al.* AI chatbots not yet ready for clinical use. *Front. Digit. Health* **5**, 60 (2023).
51. Knitza, J. *et al.* Accuracy and usability of a diagnostic decision support system in the diagnosis of three representative rheumatic diseases: A randomized controlled trial among medical students. *Arthritis Res. Ther.* **23**, 233 (2021).
52. Ray, P. P. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet Things Cyber Phys. Syst.* **3**, 121–154 (2023).
53. European Parliament. Proposal for a regulation of the European Parliament and of the Council on harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts. Preprint at https://www.europarl.europa.eu/news/es/press-room/20230505IPR84904/ai-act-a-step-closer-to-the-first-rules-on-artificial-intelligence (2023).
54. Uz, C. & Umay, E. "Dr ChatGPT": Is it a reliable and useful source for common rheumatic diseases? *Int. J. Rheum. Dis.* **26**, 1343–1349 (2023).
55. Feinstein, A. R. & Cicchetti, D. V. High agreement but low Kappa: I. The problems of two paradoxes. *J. Clin. Epidemiol.* **43**, 543–549 (1990).
56. Cicchetti, D. V. & Feinstein, A. R. High agreement but low kappa: II. Resolving the paradoxes. *J. Clin. Epidemiol.* **43**, 551–558 (1990).
57. Quarfoot, D. & Levine, R. A. How robust are multirater interrater reliability indices to changes in frequency distribution? *Am. Stat.* **70**, 373–384 (2016).
58. Seghier, M. L. ChatGPT: Not all languages are equal. *Nature* **615**, 216 (2023).

Author contributions

A.M.-G.: Conceptualization of this study, methodology, review, writing (original draft preparation). Z.R.-R.: Evaluation. D.F.-N.: Evaluation. I.P.-S.: Evaluation. E.P.-C.: Evaluation. C.P.-R.: Evaluation. L.C.-O.: Evaluation. L.L.-M.: Methodology. L.A.-A.: Methodology. B.F.-G.: Conceptualization of this study. L.R.-R.: Conceptualization of this study, methodology, review. All of the authors were involved in the drafting and/or revising of the manuscript. During the preparation of this work, the author(s) used ChatGPT May 12th version, 2023, (OpenAI, San Francisco, CA, USA) as a writing aid in the composition of this scientific article, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Funding

This work was supported by the Instituto de Salud Carlos III, Ministry of Health, Madrid, Spain [RD21/002/0001]. The sponsor or funding organization had no role in the design or conduct of this research. The journal's fee was funded by the institution employing the senior author of the manuscript (Fundación Biomédica del Hospital Clínico San Carlos).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-49483-6.

Correspondence and requests for materials should be addressed to A.M.-G.