BMC Infectious Diseases

RESEARCH

Open Access

Language discrepancies in the performance of generative artificial intelligence models: an examination of infectious disease queries in English and Arabic

Malik Sallam[1,2,7*], Kholoud Al-Mahzoum[3], Omaima Alshuaib[3], Hawajer Alhajri[3], Fatmah Alotaibi[3], Dalal Alkhurainej[3], Mohammad Yahya Al-Balwah[3], Muna Barakat[4,5] and Jan Egger[6]

Abstract

Background Assessment of artificial intelligence (AI)-based models across languages is crucial to ensure equitable access and accuracy of information in multilingual contexts. This study aimed to compare AI model efficiency in English and Arabic for infectious disease queries.

Methods The study employed the METRICS checklist for the design and reporting of AI-based studies in healthcare. The AI models tested included ChatGPT-3.5, ChatGPT-4, Bing, and Bard. The queries comprised 15 questions on HIV/AIDS, tuberculosis, malaria, COVID-19, and influenza. The AI-generated content was assessed by two bilingual experts using the validated CLEAR tool.

Results In comparing AI models' performance in English and Arabic for infectious disease queries, variability was noted. English queries showed consistently superior performance, with Bard leading, followed by Bing, ChatGPT-4, and ChatGPT-3.5 (*P* = .012). The same trend was observed in Arabic, albeit without statistical significance (*P* = .082). Stratified analysis revealed higher scores for English in most CLEAR components, notably in completeness, accuracy, appropriateness, and relevance, especially with ChatGPT-3.5 and Bard. Across the five infectious disease topics, English outperformed Arabic, except for flu queries in Bing and Bard. The four AI models' performance in English was rated as "excellent", significantly outperforming their "above-average" Arabic counterparts (*P* = .002).

Conclusions Disparity in AI model performance was noticed between English and Arabic in response to infectious disease queries. This language variation can negatively impact the quality of health content delivered by AI models among native speakers of Arabic. This issue is recommended to be addressed by AI developers, with the ultimate goal of enhancing health outcomes.

Keywords AI chatbots, Infectious diseases, Language performance, Healthcare technology, Digital health queries

*Correspondence:
Malik Sallam
malik.sallam@ju.edu.jo

Full list of author information is available at the end of the article

Background

Arabic is a culturally diverse language spoken daily by over 400 million people [1]. Consequently, the Arabic language is considered an important medium for delivering health-related information to a substantial number of native speakers [2]. The pursuit of ensuring access to accurate health information in the native language is essential for effective communication and better health outcomes [3, 4].

From a global health perspective, the "big three" infectious diseases — malaria, tuberculosis (TB), and human immunodeficiency virus/acquired immunodeficiency syndrome (HIV/AIDS) — rise as prevalent health concerns [5, 6]. Additionally, the profound impact of the coronavirus disease 2019 (COVID-19) pandemic, highlighted the need for effective health communication [7]. Furthermore, influenza continues to pose significant public and global health risks with the potential to cause epidemics and pandemics; therefore, effective public health measures are needed to address influenza threats [8].

In the current digital era, lay individuals increasingly seek health information via various online platforms [9]. While these online channels — including the recent rise of artificial intelligence (AI)-based chatbots — offer convenient access to data, these digital channels also present significant challenges and concerns about the reliability of the information provided [10–12]. The prevalence of misinformation or even disinformation on these platforms can pose significant risks. Lay individuals may encounter and act upon inaccurate health-related content, potentially compromising their health and well-being [13–15].

ChatGPT (by OpenAI, San Francisco, CA), Bing (by Microsoft Corporation, Redmond, WA), and Bard (by Google, Mountain View, CA) are AI-based conversational models that emerged as promising tools for various purposes including the ability to facilitate the acquisition of health information [16–18]. These chatbots garnered notable user attention due to their ease of use and perceived effectiveness in delivering a broad spectrum of information [19]. This includes health-related content and self-diagnosis options, marking a significant advance in digital health communication and information accessibility [20–23]. Consequently, a notable surge in research interest regarding the utility of generative AI models in healthcare has been noticed [24, 25]. This interest was motivated by generative AI models' ability to synthesize and analyze huge medical data rapidly, offering possibilities for personalized medicine and enhanced diagnostic accuracy [26–30]. Studies have focused on evaluating AI effectiveness in tasks such as generating patient education materials, simulating physician-patient interactions, and automating parts of the diagnostic process [16, 31–39].

The impact of linguistics on the evolution and efficacy of Large Language Models (LLMs) is profound [40, 41]. To enhance the accessibility of LLMs across various cultural and linguistic contexts, linguistic insights are important for the development and evolution of LLMs capable of competent performance across multiple languages and dialects [42]. In healthcare, multilingual LLMs can equate access to medical information and healthcare services through circumventing language barriers [16, 27, 43]. Patients and healthcare professionals who speak different languages can benefit from real-time translation services, ensuring that crucial health information is both accessible and understandable to diverse populations [44]. Such advancements are important for the successful integration of AI technologies into healthcare, improving operational efficiency and enhancing patient care [27]. The application of deep learning and AI within the healthcare sector has led to transformative developments. Examples include the development of models capable of accurate differentiation between cancerous versus normal blood cells, determining the severity of COVID-19 through the analysis of radiographic images, and improving the accuracy of malaria parasite detection in blood samples [45–47].

While generative AI-based models like ChatGPT, Bing, and Bard are promising in disseminating health information and improving health literacy, it is crucial to recognize their limitations [16, 48]. For example, a notable issue is the occurrence of "hallucinations" where AI models generate plausible but incorrect responses [49]. This is particularly concerning in the context of health-related information, where such inaccuracies could lead to severe negative consequences [50]. Understanding and addressing the limitations of AI-based models is essential for the safe and effective use of AI in healthcare communication [16, 33, 48, 51].

The performance of generative AI-based models is highly influenced by the quality of the underlying training data [52]. Therefore, variations in AI-based model performance would reasonably be anticipated across different languages and cultural contexts [53]. Consequently, a thorough assessment of AI-based model performance in a variety of languages is needed, to ensure the accuracy and reliability of these models in diverse languages.

To address this critical issue, this study aimed to evaluate the performance of a group of popular AI-based models, namely ChatGPT, Bing, and Bard in English and Arabic languages. The focus of this study involved one aspect of health-related information by choosing queries on five infectious diseases (HIV/AIDS, TB, malaria, COVID-19, and influenza). By exploring the capabilities and shortcomings of these AI-based models in the context of health information dissemination in Arabic, the study aimed to highlight the need to enhance the quality of healthcare content that would be provided to native speakers of Arabic for better health outcomes within Arab communities. Additionally, the study sought to identify potential disparities in the language performance of AI-based models, which are predominantly trained on English datasets.

The evaluation of generative AI-based models across Arabic and English languages, particularly within the context of infectious diseases, holds specific importance for the following reasons. For example, infectious diseases remain a global health concern, necessitating rapid communication and dissemination of accurate information, which was manifested during the COVID-19 pandemic [54]. The deployment of AI models can facilitate immediate health guidance and insights, with an important need for consistent performance across languages, to ensure effective public health communication [55, 56]. Additionally, variability in the performance of generative AI models across languages may create disparities in access to accurate and dependable health information [57]. Such disparities have the potential to amplify health inequities. Therefore, evaluating generative AI models' applicability in handling queries related to infectious diseases across different linguistic contexts is essential to identify and address potential deficiencies to ensure equitable access to health information across the globe.

Methods

Study design

This descriptive study was designed following the METRICS checklist for AI-based studies in healthcare [58, 59]. This framework involves careful consideration of the features and settings of the AI models, a detailed evaluation methodology, and clear specifications of prompts, languages, and data sources. Additionally, the study rigorously addressed factors such as the count of queries, the individual factors in query selection, and the subjectivity inherent in evaluation of the generated content. The study design also considered the issues of randomization and the range of topics tested, adhering to the principles of transparency and thoroughness.

Ethics statement

This study was approved by the Institutional Review Board (IRB) at the Faculty of Pharmacy, Applied Science Private University (Approval number: 2023-PHA-51, on 23 November 2023).

Features of the AI models tested

Four AI-based models were employed in this study as follows. Two versions of ChatGPT (the publicly available GPT-3.5 and the more advanced, subscription-based GPT-4), Microsoft Bing, using the more balanced conversational style, and Google Bard Experiment, both available for free. To ensure content replicability, each model was tested under its default configuration. The prompting of these AI models was carried out concurrently on a single day by the first author (M.S.), specifically on 23 December 2023, to maintain consistency and control for time-sensitive variables in their performance assessment.

Features and count of the queries used to test the AI models

In this study, 15 distinct queries were executed on each AI model. This query count was based on the calculated sample size necessary for comparing means between two groups: $n = (Z_{\alpha/2}+Z_{\beta})^2 *2*\sigma^2 / d^2$ considering a 90% confidence level, an 80% desired power, and an assumed difference and variance of 1 [60]. This yielded a minimum of 13 queries to elucidate possible differences between the two languages effectively. This decision was guided by the aim to effectively examine the AI-generated responses, while also accommodating the operational constraints imposed by the rate limits of the AI models.

Sources of data to formulate the infectious disease queries

The queries purposefully examined five common infectious diseases, focusing on transmission, treatment, diagnosis, prevention, and epidemiology. For each disease, three queries were randomly selected using Excel's randomize function from a pool of 15 questions per topic to minimize selection bias. The initial pool of queries were retrieved from credible English sources and covered key questions on HIV/AIDS, malaria, TB, COVID-19, and influenza as follows [61–70]. For HIV/AIDS, the three questions were: (1) What is the extent of risk of HIV transmission through French kiss? (2) What is the extent of risk of HIV transmission through *hijama*? (3) Why gays have higher chance of getting HIV infection? For malaria, the three questions were (1) Is malaria a contagious disease? (2) Is it considered safe for me to breastfeed while taking an antimalarial drug? (3) How do I know if I have malaria for sure? For TB, the three questions were: (1) Who doesn't get sick from tuberculosis? (2) How can TB be tested for? (3) Is BCG vaccination recommended for all children? For COVID-19, the three questions were: (1) Can COVID-19 be passed through breastfeeding? (2) Can COVID-19 infection affect HIV test result? (3) What is long COVID-19 condition? Finally, for influenza, the three questions were: (1) Can I get a COVID-19 vaccine and flu vaccine at the same visit? (2) Is it possible to have both COVID-19 and flu at the same time? (3) When will flu activity begin and when will it peak?

The questions were translated into Arabic by one bilingual author (M.B.) and back translated by another (M.S.), with subsequent discussions among the two authors leading to minor modifications for clarity.

Specificity of prompts used

The prompting approach for each AI model involved using the prompts as exact questions without any feedback. This was ensured by selecting the "New Chat" or "New Topic" options for each query. The "Regenerate Response" feature was not utilized to maintain the integrity of first responses. Additionally, each query was initiated as a new chat or topic when switching languages to prevent any carryover effects between languages. This approach was critical to ensure that the responses for the same query in different languages were independent and not influenced by previous interactions.

Evaluation of the AI generated content

The evaluation of the AI-generated content was conducted independently by two authors with expertise in infectious disease from clinical microbiology (M.S.) and pharmacy (M.B.) perspectives. To minimize subjectivity in the evaluation process, a consensus key response was formulated prior to assessment based on the query sources. The evaluation was based on the CLEAR tool across 5 components as follows: Completeness, Lack of false information (accuracy), Evidence-based content, Appropriateness, and Relevance [71]. Each component was assessed using a 5-point Likert scale ranging from 5 (excellent) to 1 (poor).

Statistical and data analyses

Statistical analyses were conducted using IBM SPSS Statistics for Windows, Version 26, with a significance level set at $P<.050$. The average CLEAR scores across the two raters were utilized, including both component-specific and overall CLEAR scores. Based on the non-normal distribution of the scale variables assessed using the Shapiro-Wilk test, the Kruskal Wallis H (K-W) and Mann Whitney *U* (M-W) tests were used for mean difference testing. The overall CLEAR scores were categorized for descriptive analysis of content quality as follows: 1–1.79 as "poor", 1.80–2.59 as "below average", 2.60–3.39 as "average", 3.40–4.19 as "above average", and 4.20–5.00 as "excellent".

To assess the consistency of evaluation between the two raters, we employed Intraclass Correlation Coefficient (ICC) average measures with two-way mixed effects to quantify inter-rater agreement. The inter-rater reliability analysis was conducted on a set of 120 responses, evenly split between English and Arabic, with 60 responses per language. The disagreement among the two raters was not resolved through post-hoc discussions after the evaluations were conducted by the raters. This decision was based on a deliberate methodological choice to maintain the objectivity of the initial independent assessments.

Results

Overall performance of each AI model in English vs. Arabic

Using the average CLEAR scores, variability was observed between the content generated in English based on the model with the best performance for Bard (mean CLEAR: 4.6±0.68) followed by Bing (mean CLEAR: 4.37±0.59), ChatGPT-4 (mean CLEAR: 4.36±0.76), and ChatGPT-3.5 (mean CLEAR: 4.15±0.68, $P=.012$, K-W). In Arabic, the same differences were observed; nevertheless, the differences lacked statistical significance (mean CLEAR: 4.39±0.89 for Bard, 4.21±0.72 for Bing, 4.13±0.97 for ChatGPT-4, and 3.81±0.68 for ChatGPT-3.5, $P=.082$, K-W).

Consistent superior performance of the four AI models tested was noted in English queries as opposed to the Arabic content (Table 1). However, statistically significant differences were observed only with ChatGPT-3.5 and Bard. Based on the descriptive assignments of the CLEAR scores, the four AI models content in English was described as "Excellent" while the performance of

Table 1 The performance of the four AI models tested in English and arabic stratified per average CLEAR scores

AI [1] model	Language	C [3]	L [4]	E [5]	A [6]	*R*[7]	CLEAR [8]
ChatGPT-3.5	English	4.5±0.5	4.53±0.81	4.37±0.64	4.63±0.4	4.43±0.42	4.49±0.49
	Arabic	3.8±0.8	3.87±1.01	3.83±1.06	3.63±0.93	3.9±0.83	3.81±0.68
P value [2]		**0.012**	**0.045**	0.074	**0.002**	**0.006**	**0.010**
ChatGPT-4	English	4.6±0.39	4.73±0.56	4.6±0.39	4.67±0.41	4.4±0.47	4.6±0.37
	Arabic	4.2±0.98	4.17±1.26	4.03±1.26	4.1±0.97	4.13±0.85	4.13±0.97
P value		0.292	0.126	0.519	**0.036**	0.345	0.144
Bing	English	4.23±0.65	4.67±0.52	4.77±0.32	4.43±0.59	4.57±0.32	4.53±0.39
	Arabic	3.9±0.93	4.07±1.28	4.27±0.98	4.2±0.77	4.63±0.44	4.21±0.72
P value		0.318	0.256	0.159	0.450	0.398	0.381
Bard	English	4.73±0.42	4.97±0.13	4.87±0.4	4.93±0.26	4.57±0.37	4.81±0.25
	Arabic	4.53±0.9	4.13±1.23	4.33±1.03	4.67±0.9	4.3±0.8	4.39±0.89
P value		0.639	**0.011**	0.082	0.169	0.322	**0.049**

[1]AI: Artificial intelligence; [2]*P* value: Calculated using the Kruskal Wallis test; [3]C: Completeness; [4]L: Lack of false information; [5]E: Evidence-based; [6]A: Appropriateness; [7]R: Relevance; [8]CLEAR: The average CLEAR scores based on the scoring of two independent raters. The significant *P* values are highlighted in bold style

both ChatGPT models in Arabic was "above average", as opposed to "excellent" performance in Arabic in Bing and Bard.

Performance of each AI model stratified per CLEAR components

In stratified analysis of AI model performance across the five CLEAR components, English content consistently scored higher in 19 out of 20 comparisons (95%). The exception was Bing's superior relevance score in Arabic compared to English. Statistically significant differences were observed with ChatGPT-3.5 and Bard. Specifically, ChatGPT-3.5 exhibited superior performance in completeness and relevance in English as opposed to Arabic content, while both ChatGPT-3.5 and Bard showed higher accuracy (lack of false information) in English. Additionally, ChatGPT-3.5 and ChatGPT-4 content in English outperformed the Arabic content in appropriateness (Table 1).

Upon evaluation of the CLEAR tool across English and Arabic, it was observed that inter-rater agreement varied across different CLEAR components (Table 2). English evaluations demonstrated lower ICC values, particularly in Completeness and Relevance, which suggests a considerable disagreement between the two raters. This contrast was less pronounced in CLEAR components assessing factual accuracy, such as Lack of False Information (Table 2). For Arabic evaluations, the agreements between the two raters were consistently higher across all CLEAR components, with less pronounced differences in raters' assessments (Table 2).

Table 2 Intraclass correlation coefficients (ICC) for English and Arabic evaluations by the two expert raters stratified per CLEAR components

CLEAR component	Language	Intraclass Correlation	
		Average Measures	*P* value
Completeness	English	0.384	**0.033**
	Arabic	0.841	**< 0.001**
Lack of false information	English	0.758	**< 0.001**
	Arabic	0.932	**< 0.001**
Evidence-based	English	0.518	**0.003**
	Arabic	0.859	**< 0.001**
Appropriateness	English	0.499	**0.004**
	Arabic	0.841	**< 0.001**
Relevance	English	−0.234	0.789
	Arabic	0.723	**< 0.001**
Overall CLEAR	English	0.714	**< 0.001**
	Arabic	0.912	**< 0.001**

The significant *P* values are highlighted in bold style

Performance of each AI model stratified per infectious disease topic

Out of the 20 comparisons across the 2 languages for the four AI models, higher average CLEAR scores were observed across all infectious disease topics in English content, with the exception of better performance in Arabic for the influenza queries in Bing and Bard (Fig. 1).

In English, Bard topped the performance in HIV/AIDS, malaria, TB, and COVID-19 while ChatGPT-3.5 topped the performance in influenza. The lowest level of performance for HIV/AIDS and COVID-19 was seen in ChatGPT-3.5 content and for malaria and TB, the lowest performance in English was seen with Bing content, while the lowest for influenza was in Bard (Fig. 2A).

In Arabic, Bard also topped the performance in four topics (TB, COVID-19, influenza, and malaria together with ChatGPT-4), while the best performance for HIV/AIDS was observed for Bing. The lowest level of performance per topic in Arabic was seen for ChatGPT-3.5 in HIV/AIDS, malaria, and COVID-19, and the lowest for TB was the Arabic content of Bing and the lowest for influenza was content generated by ChatGPT-4 (Fig. 2B).

Descriptive labeling of the performance of each AI model in English vs. Arabic

Compiled together as shown in (Fig. 3), the overall performance of the four models in English was "excellent" with a mean CLEAR score of 4.6 ± 0.4 while in Arabic it was "above average" with a mean CLEAR score of 4.1 ± 0.82 ($P = .002$, M-W).

Discussion

In this study, we investigated one crucial aspect of generative AI models' utility in acquisition of health information. This involves testing the hypothesis of existent language disparity in generative AI model performance. Specifically, the study pursuit was in the context of infectious diseases which represent a significant global health burden. Such a quest appears timely and relevant as generative AI models are increasingly accessed by lay individuals for health information [21, 72]. Concerns emerged regarding the potential of generative AI models to produce harmful or misleading content with recurring calls for ethical guidance, benchmarking, and human oversight [73–76].

The key finding in this study was the overall lower performance of the tested AI models in Arabic compared to English. In this study, the overall Arabic performance of generative AI models in the context of infectious disease queries could be labeled as "above average" as opposed to "excellent" performance in English. Additionally, the differences in performance across the two languages showed statistical significance in ChatGPT-3.5 and Bard. Another important observation was the uniformly

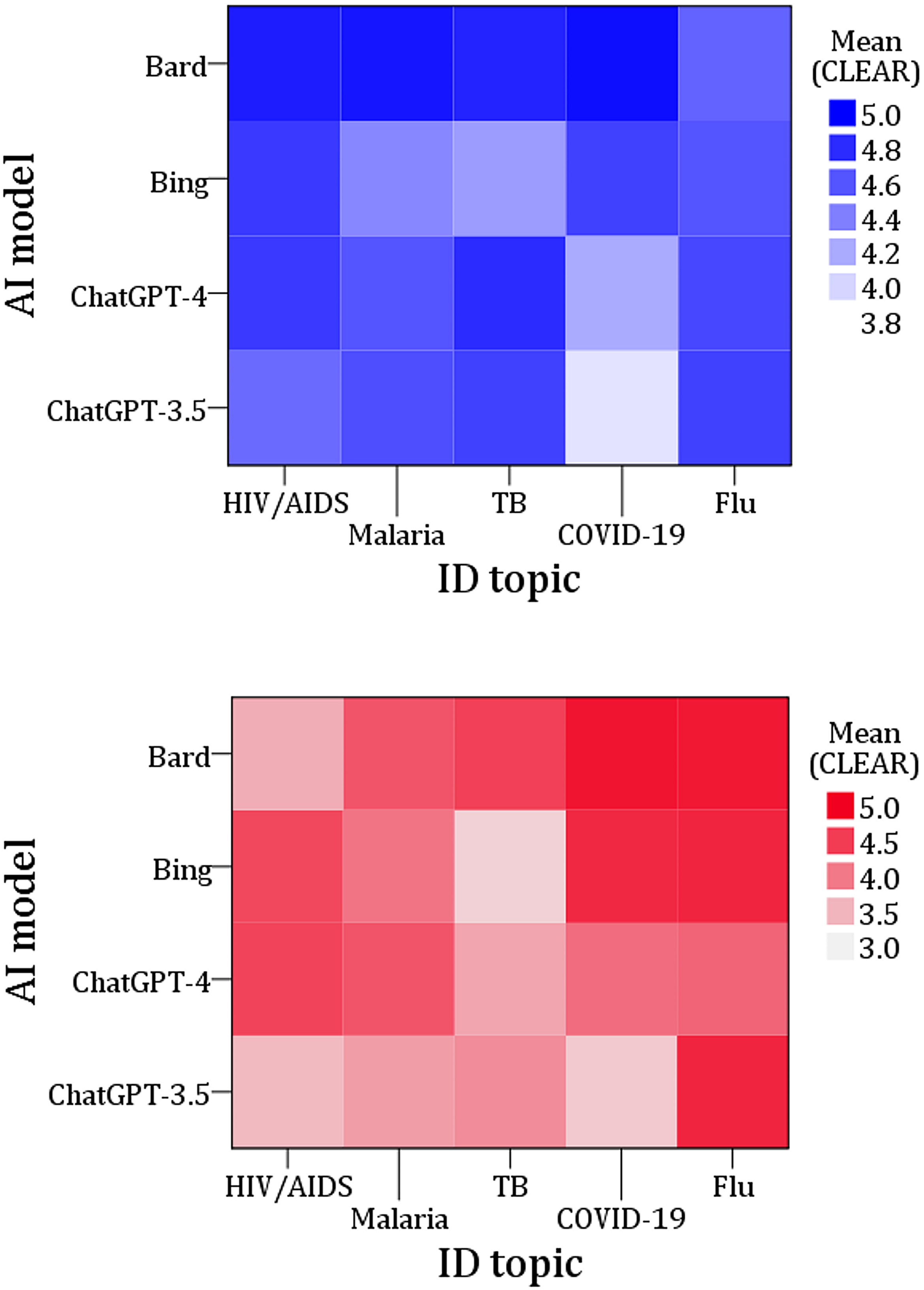

Fig. 1 Heat maps of the four artificial intelligence models' performance in English (blue) and Arabic (red) based on infectious disease queries. Assessment was based on the average CLEAR scores. COVID-19: Coronavirus disease 2019; TB: Tuberculosis; HIV/AIDS: Human immunodeficiency virus/ acquired immunodeficiency syndrome

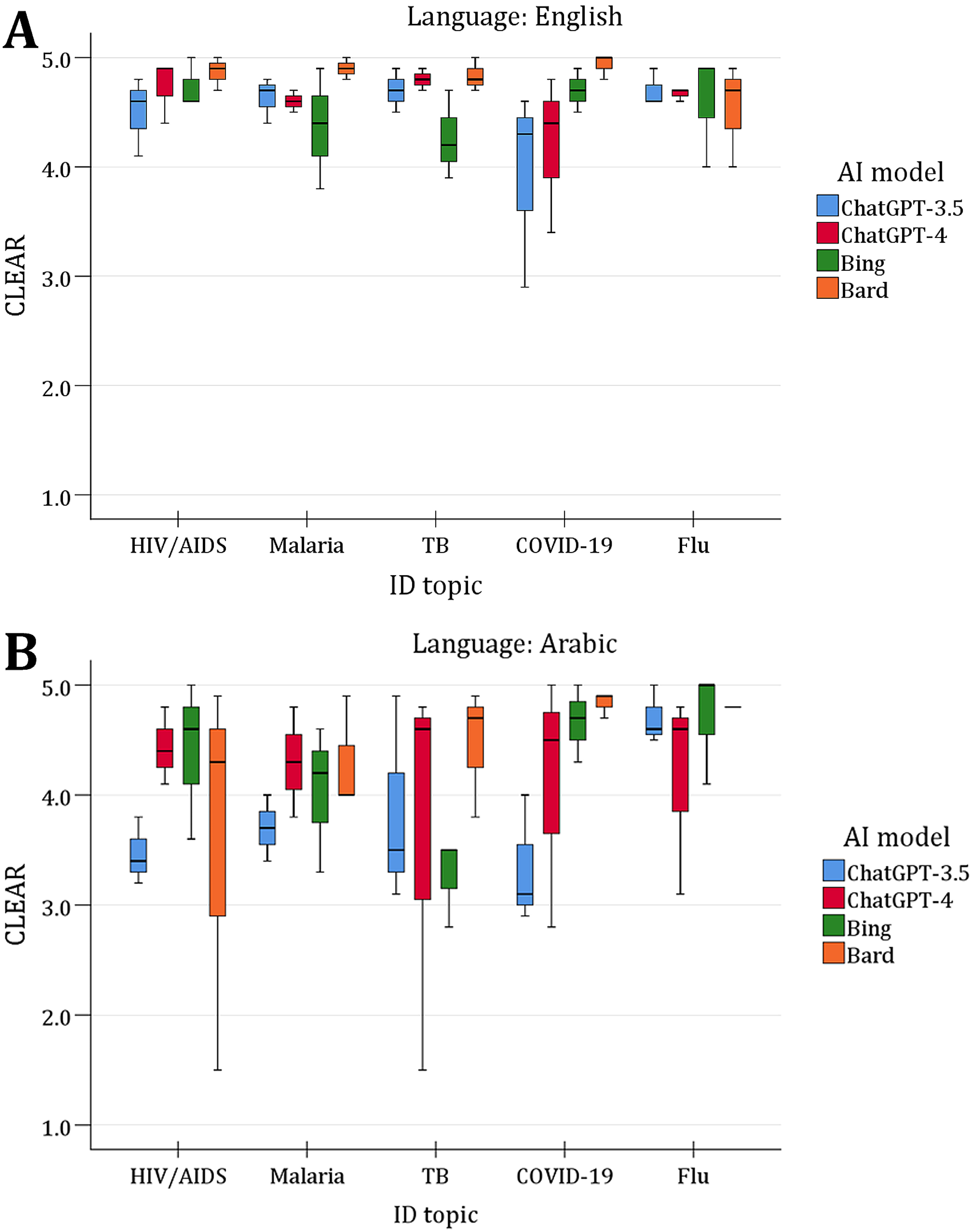

Fig. 2 Box plots of the four artificial intelligence models' performance in English (A) and Arabic (B) showing variability in CLEAR scores. Assessment was based on the average CLEAR scores. COVID-19: Coronavirus disease 2019; TB: Tuberculosis; HIV/AIDS: Human immunodeficiency virus/acquired immunodeficiency syndrome

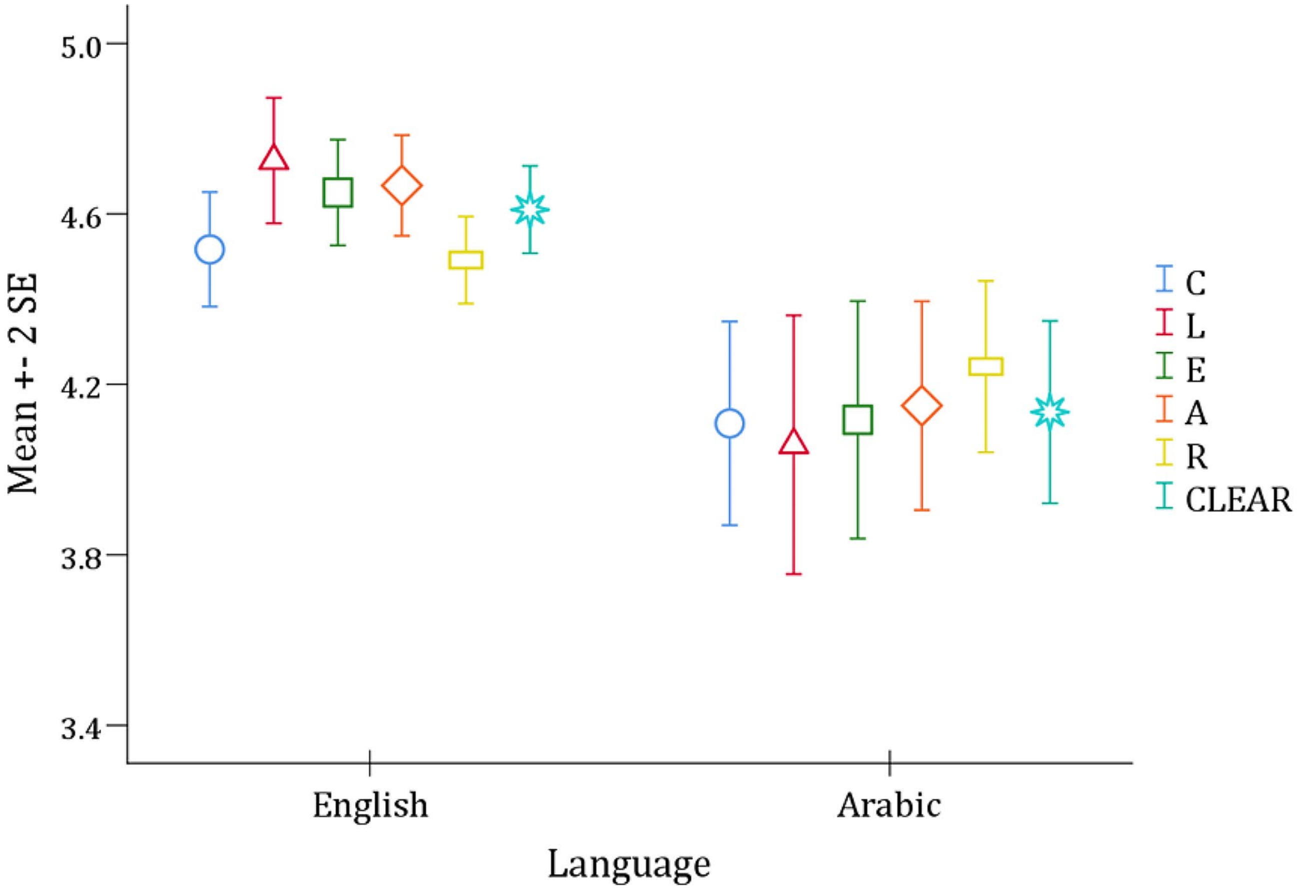

Fig. 3 Error bars of the four artificial intelligence models' performance compiled together and showing the five CLEAR components and the overall CLEAR scores stratified per language. SE: Standard error of the mean; C: Completeness; L: Lack of false information; E: Evidence support; A: Appropriateness; R: Relevance

excellent performance of the four generative AI models in English. This consistency highlights the effectiveness of these models in the English language in the context of infectious disease queries. Additionally, a consistent pattern where the four AI models exhibited superior performance in English extended across all the five tested infectious disease topics. However, a notable variability in performance in Arabic was evident, particularly in handling topics related to HIV/AIDS, TB, and COVID-19.

The disparity in generative AI model performance across languages may be attributed to the varying qualities of the AI training datasets [77]. Prior research that sought to characterize such disparity in generative AI model performance across languages remains limited with variable results despite its timeliness and significance [78–83]. This includes even fewer studies that compared the AI content generated for the same queries in multiple languages [84].

Several studies assessed AI model performance in non-English languages with variable results despite the overall trend of below-bar performance in non-English languages. For example, Taira et al. tested ChatGPT performance in the Japanese National Nursing Examination in the Japanese language in five consecutive years [85]. Despite approaching the passing threshold in four years and passing the 2019 exam, the results indicated the relative weakness of ChatGPT in Japanese [85]. Nevertheless, attributing this result to language limitations alone is challenging, given the superior performance of ChatGPT-4 in the Japanese language compared to medical residents in the Japanese General Medicine In-Training Examination, as reported by Watari et al. [86]. This study also exposed ChatGPT-4 limitations in test aspects requiring empathy, professionalism, and contextual understanding [86]. Conversely, another recent study highlighted ChatGPT-4's capabilities in acting in human-like behavior, being helpful and demonstrating empathy, which suggests variability in AI performance based on the nature of the task required [87]. These contrasting findings highlight the need for further studies to explore the emotional intelligence aspects of generative AI models.

In a study by Guigue et al., ChatGPT limitations in French were evident, with only one-third of questions correctly answered in a French medical school entrance examination, mirroring its performance in obstetrics and gynecology exam [88]. Additionally, the worse performance of ChatGPT compared to students was seen in the context of family medicine questions in the Dutch language [89]. Conversely, in the Polish Medical Final Examination, ChatGPT demonstrated similar effectiveness in both English and Polish, with a marginally higher accuracy in English for ChatGPT-3.5 [90]. In Portuguese, ChatGPT-4 displayed satisfactory results in the 2022 Brazilian National Examination for Medical Degree Revalidation [91].

In the context of the Arabic language and in line with our findings, Samaan et al. showed less accurate performance of ChatGPT in Arabic compared to English in cirrhosis-related questions [92]. In a non-medical context, Banimelhem and Amayreh showed that ChatGPT's performance as an English-to-Arabic machine translation tool was suboptimal [93]. In a comprehensive study, Khondaker et al. revealed that smaller, Arabic-fine-tuned models consistently outperformed ChatGPT, indicating significant room for improvement in multilingual capabilities, particularly in Arabic dialects [94]. In the current study, our results suggested that the pattern of lower performance in Arabic extends to all tested AI models despite lacking significance in Bing and Bard.

The use of the CLEAR tool in this study was crucial for pinpointing specific areas for improvement in each language. Specifically, the study findings revealed that in both GPT-3.5 and GPT-4 models, the appropriateness in Arabic lagged behind English. This highlights key areas for enhancement in Arabic, such as the need to improve areas of ambiguities in the generated content and the need to organize the content in a more effective style. Additionally, accuracy issues observed in ChatGPT-3.5 and Bard highlighted the need for content verification particularly in health-related queries as well as the necessity of acknowledging the potential for inaccuracies in these models (e.g., through clear flagging of potential inaccuracies within the generated responses). The enhanced performance of ChatGPT-4 in both English and Arabic, relative to its predecessor GPT-3.5, has been demonstrated in previous research, which highlights the rapid significant advancements in generative AI models [95]. These improvements are attributed to refined training algorithms and larger, more diverse datasets, which enable the AI models to generate more accurate and contextually appropriate responses [52].

In light of this study findings, several recommendations for subsequent research could be outlined to enhance the applicability of generative AI models in healthcare as follows. First, the AI developers are recommended to integrate cultural and linguistic diversity aspects into the generative AI models, especially for AI algorithms aimed generate health-related content. Addressing the linguistic disparities revealed in this study is important to enhance equitable access to health information across diverse languages and cultural contexts. Second, further research is needed to confirm if the observed discrepancies in generative AI models' performance extend beyond the English and Arabic languages examined in this study. Thus, expanding the scope of research to include a wider array of languages and dialects can improve the collective understanding of linguistic biases inherent in generative AI models [96]. This is particularly relevant for languages underrepresented in the current AI training datasets.

Moreover, the ethical and cultural consequences of deploying AI in healthcare necessitate rigorous scrutiny [97]. Real-world implementation studies of generative AI models in healthcare across different linguistic regions could shed light on the practical limitations of implementing generative AI tools in patient care [27, 98]. Incorporating feedback from non-English speakers into the AI development process can help to identify unique user needs and preferences, which would help to guide the development of more accessible and user-friendly AI algorithms.

Finally, the establishment of rigorous standards and guidelines for the development and assessment of multilingual AI models in healthcare is important [99]. Such standards can be helpful to ensure that the generative AI tools meet the required standards prior to deployment in healthcare. A collaborative effort among AI developers, researchers, and healthcare professionals is essential to ensure the applicability of generative AI models in disseminating accurate healthcare information tailored to different cultural and geographic settings [16, 27].

Lastly, it is important to interpret the results of the study in light of several limitations as follows: First, the limited number of queries tested on each model, albeit sufficient to reveal potential disparities might limit the generalizability of the findings. Future studies can benefit from incorporating a larger and more diverse set of queries to further validate and refine the findings of this study. Second, the assignment of CLEAR scores may vary if assessed by different raters. To mitigate potential measurement bias, this study employed key answers derived from credible sources as an objective benchmark before CLEAR scoring of the AI generated content. Third, the study did not account for the various Arabic dialects, focusing only on the Standard Arabic. Future research could expand on this particular issue in light of the previous evidence showing potential variability in dialectical performance [94, 100]. Fourth, in this study, we adhered to pre-established consensus key responses to maintain objectivity upon expert assessment of the AI-generated content. However, this approach limited our ability to capture dynamic consensus that could emerge from direct raters' interactions. The observed discrepancies in expert raters' agreement, especially in the evaluations of AI-generated content in English, suggest that linguistic complexity and the subjective nature of certain CLEAR components impacted the consistency of assessments. The notably low agreement on the Relevance of content in English might reflect broader issues in interpreting relevance across different medical contexts, where the raters may hold divergent views based on their backgrounds and expertise. Future studies could benefit from refined guidelines for these components, potentially incorporating more structured and detailed criteria to aid raters in achieving higher consistency. Fifth, in translating the queries from English to Arabic, the study employed a simplified practical approach where two bilingual authors independently translated the queries. This expedient method did not follow the rigorous, standardized procedures recommended for cross-cultural healthcare research, such as those outlined by Sousa and Rojjanasrirat [101]. Consequently, this might have introduced variations in the semantic equivalence of the queries across languages, potentially affecting the reliability and validity of the responses. Future research should consider implementing a more structured translation methodology, including validation by a panel of linguistic and subject matter experts. Finally, future studies can benefit from including a broader range of queries involving not only infectious disease topics to achieve a more comprehensive understanding of AI performance in diverse health and linguistic contexts. Addressing these limitations in future studies can help to advance the collective understanding of multilingual generative AI applications and to enhance the generative AI tools' reliability and equity in global healthcare settings.

Conclusions

This study demonstrated the language discrepancy in generative AI models' performance. Specifically, a generally inferior performance of the tested generative AI models in Arabic was observed compared to English, despite being rated "above average". These findings highlight the language-based performance gaps in commonly used generative AI chatbots. This suggests the need for enhancements in AI performance in Arabic. Nevertheless, further research is needed across various health topics and utilizing different languages to discern this pattern. To achieve equitable global health standards, it is important to consider the cultural and linguistic diversity in generative AI model fine-tuning for widespread applicability.

Abbreviations

AI	Artificial intelligence
CLEAR	Completeness, Lack of false information, Evidence, Appropriateness, and Relevance
COVID-19	Coronavirus disease 2019
HIV/AIDS	Human immunodeficiency virus/acquired immunodeficiency syndrome
ICC	Intraclass Correlation Coefficient
K-W	Kruskal Wallis H test
LLM	Large Language Model
METRICS	Model, Evaluation, Timing, Range/Randomization, Individual, Count, and Specificity of prompt and language
M-W	Mann Whitney U test
TB	Tuberculosis

Author contributions

Conceptualization: M.S.; Data curation: M.S., K.A.-M., O.A., H.A., F.A., D.A., M.Y.A.-B., M.B., J.E.; Formal analysis: M.S.; Investigation: M.B. and M.S.; Methodology: K M.S., K.A.-M., O.A., H.A., F.A., D.A., M.Y.A.-B., M.B., J.E.; Visualization: M.S.; Project administration: M.S.; Supervision: M.S., J.E; Writing - original draft: M.S.; Writing - review & editing: M.S., K.A.-M., O.A., H.A., F.A., D.A., M.Y.A.-B., M.B., J.E.; All authors contributed to the article and approved the submitted version.

Funding

This research received no funding.

Data availability

The datasets analyzed during the current study are available in the public data tool Open Science Framework (OSF), using the following link: https://osf.io/f86tk/; DOI: 10.17605/OSF.IO/F86TK.

Declarations

Ethics approval and consent to participate

This study was approved by the Institutional Review Board (IRB) at the Faculty of Pharmacy, Applied Science Private University (Approval number: 2023-PHA-51, on 23 November 2023). The consent to participate in this study is not applicable based on no involvement of human subjects.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]Department of Pathology, Microbiology and Forensic Medicine, School of Medicine, The University of Jordan, Amman 11942, Jordan
[2]Department of Translational Medicine, Faculty of Medicine, Lund University, Malmö 22184, Sweden
[3]School of Medicine, The University of Jordan, Amman 11942, Jordan
[4]Department of Clinical Pharmacy and Therapeutics, Faculty of Pharmacy, Applied Science Private University, Amman 11931, Jordan
[5]MEU Research Unit, Middle East University, Amman 11831, Jordan
[6]Institute for AI in Medicine (IKIM), University Medicine Essen (AöR), Essen, Germany
[7]Department of Clinical Laboratories and Forensic Medicine, Jordan University Hospital, Queen Rania Al-Abdullah Street-Aljubeiha, P.O. Box: 13046, Amman, Jordan

Received: 2 January 2024 / Accepted: 6 August 2024
Published online: 08 August 2024

References

1. UNESCO. World Arabic Language Day. 25. December 2023, 2023. Updated 18 December 2023. Accessed 25 December 2023, 2023. https://www.unesco.org/en/world-arabic-language-day
2. Alfakhry GM, Dashash M, Jamous I. Native Arabic Language Use Acceptability and Adequacy in Health Professional Instruction: Students and Faculty's Perspectives. *Health Professions Education*. 2020/12/01/, 2020;6(4):454–464. doi:10.1016/j.hpe.2020.06.004.

3. Al Shamsi H, Almutairi AG, Al Mashrafi S, Al Kalbani T. Implications of Language Barriers for Healthcare: a systematic review. Oman Med J Mar. 2020;35(2):e122. https://doi.org/10.5001/omj.2020.40.
4. Gazzaz ZJ, Baig M, Albarakati M, Alfalig HA, Jameel T. Language barriers in understanding Healthcare Information: arabic-speaking students' comprehension of Diabetic questionnaires in Arabic and English languages. Cureus Oct. 2023;15(10):e46777. https://doi.org/10.7759/cureus.46777.
5. Makam P, Matsa R. Big Three infectious diseases: Tuberculosis, Malaria and HIV/AIDS. Curr Top Med Chem. 2021;21(31):2779–99. https://doi.org/10.2174/1568026621666210916170417.
6. Bhutta ZA, Sommerfeld J, Lassi ZS, Salam RA, Das JK. Global burden, distribution, and interventions for infectious diseases of poverty. *Infectious Diseases of Poverty*. 2014/07/31 2014;3(1):21. https://doi.org/10.1186/2049-9957-3-21
7. Finset A, Bosworth H, Butow P, et al. Effective health communication - a key factor in fighting the COVID-19 pandemic. Patient Educ Couns May. 2020;103(5):873–6. https://doi.org/10.1016/j.pec.2020.03.027.
8. Fauci AS. Pandemic influenza threat and preparedness. Emerg Infect Dis Jan. 2006;12(1):73–7. https://doi.org/10.3201/eid1201.050983.
9. Jia X, Pang Y, Liu LS. Online Health Information seeking behavior: a systematic review. Healthc (Basel) Dec. 2021;16(12):1740. https://doi.org/10.3390/healthcare9121740.
10. Dalmer NK. Questioning reliability assessments of health information on social media. J Med Libr Assoc Jan. 2017;105(1):61–8. https://doi.org/10.5195/jmla.2017.108.
11. Moretti FA, Oliveira VE, Silva EM. Access to health information on the internet: a public health issue? *Rev Assoc Med Bras (*1992*)*. Nov-Dec 2012;58(6):650-8. https://doi.org/10.1590/s0104-42302012000600008
12. Abdaljaleel M, Barakat M, Mahafzah A, Hallit R, Hallit S, Sallam M. TikTok Content on measles-Rubella Vaccine in Jordan: a cross-sectional study highlighting the spread of Vaccine Misinformation. Narra J. 2024;4(2):e877. https://doi.org/10.52225/narra.v4i2.877.
13. Fridman I, Johnson S, Elston Lafata J. Health Information and Misinformation: a Framework to Guide Research and Practice. JMIR Med Educ Jun. 2023;7:9:e38687. https://doi.org/10.2196/38687.
14. Suarez-Lledo V, Alvarez-Galvez J. Prevalence of Health Misinformation on Social Media: systematic review. J Med Internet Res Jan. 2021;20(1):e17187. https://doi.org/10.2196/17187.
15. Meyrowitsch DW, Jensen AK, Sørensen JB, Varga TV. AI chatbots and (mis) information in public health: impact on vulnerable communities. Front Public Health. 2023;11:1226776. https://doi.org/10.3389/fpubh.2023.1226776.
16. Sallam M. ChatGPT Utility in Healthcare Education, Research, and practice: systematic review on the promising perspectives and valid concerns. Healthc (Basel) Mar. 2023;19(6):887. https://doi.org/10.3390/healthcare11060887.
17. Sallam M, Salim NA, Al-Tammemi AB, et al. ChatGPT output regarding compulsory vaccination and COVID-19 vaccine conspiracy: a descriptive study at the outset of a paradigm shift in online search for information. Cureus Feb. 2023;15(2):e35029. https://doi.org/10.7759/cureus.35029.
18. Choudhury A, Elkefi S, Tounsi A. Exploring factors influencing user perspective of ChatGPT as a technology that assists in healthcare decision making: a cross sectional survey study. medRxiv. 2023. 2023.12.07.23299685.
19. Abdaljaleel M, Barakat M, Alsanafi M, et al. A multinational study on the factors influencing university students' attitudes and usage of ChatGPT. Sci Rep. 2024;14:1983. https://doi.org/10.1038/s41598-024-52549-8.
20. Sallam M, Salim NA, Barakat M, et al. Assessing Health students' attitudes and usage of ChatGPT in Jordan: Validation Study. JMIR Med Educ Sep. 2023;5:9:e48254. https://doi.org/10.2196/48254.
21. Shahsavar Y, Choudhury A. User intentions to Use ChatGPT for self-diagnosis and health-related Purposes: cross-sectional survey study. JMIR Hum Factors May. 2023;17:10:e47564. https://doi.org/10.2196/47564.
22. Yilmaz Muluk S, Olcucu N. Comparative Analysis of Artificial Intelligence Platforms: ChatGPT-3.5 and GoogleBard in identifying red flags of low back Pain. Cureus. 2024/7/01 2024;16(7):e63580. https://doi.org/10.7759/cureus.63580
23. Mijwil M, Abotaleb M, Guma ALI, Dhoska K. Assigning Medical professionals: ChatGPT's contributions to Medical Education and Health Prediction. Mesopotamian J Artif Intell Healthc. 2024;07/20:2024:76–83. https://doi.org/10.58496/MJAIH/2024/011.
24. Khan N, Khan Z, Koubaa A, Khan MK, Salleh R. Global insights and the impact of generative AI-ChatGPT on multidisciplinary: a systematic review and bibliometric analysis. Connection Sci. 2024;36(1):2353630. https://doi.org/10.1080/09540091.2024.2353630. 2024/12/31.
25. Sallam M. Bibliometric top ten healthcare-related ChatGPT publications in the first ChatGPT anniversary. Narra J. 2024;4(2):e917. https://doi.org/10.52225/narra.v4i2.917.
26. Ghebrehiwet I, Zaki N, Damseh R, Mohamad MS. Revolutionizing personalized medicine with generative AI: a systematic review. Artif Intell Rev. 2024/04/25 2024;57(5):128. https://doi.org/10.1007/s10462-024-10768-5
27. Sallam M, Al-Farajat A, Egger J. Envisioning the future of ChatGPT in Healthcare: insights and recommendations from a systematic identification of Influential Research and a call for Papers. Jordan Med J. 2024;02/19(1):95–108. https://doi.org/10.35516/jmj.v58i1.2285.
28. Krishnan G, Singh S, Pathania M, et al. Artificial intelligence in clinical medicine: catalyzing a sustainable global healthcare paradigm. Front Artif Intell. 2023;6:1227091. https://doi.org/10.3389/frai.2023.1227091.
29. Pressman SM, Borna S, Gomez-Cabello CA, Haider SA, Haider CR, Forte AJ. Clinical and Surgical applications of large Language models: a systematic review. J Clin Med May. 2024;22(11):3041. https://doi.org/10.3390/jcm13113041.
30. Di Sarno L, Caroselli A, Tonin G, et al. Artificial Intelligence in Pediatric Emergency Medicine: applications, challenges, and future perspectives. Biomedicines. 2024;12(6):1220. https://doi.org/10.3390/biomedicines12061220.
31. Ali O, Abdelbaki W, Shrestha A, Elbasi E, Alryalat MAA, Dwivedi YK. A systematic literature review of artificial intelligence in the healthcare sector: benefits, challenges, methodologies, and functionalities. J Innov Knowl. 2023;8(1):100333. https://doi.org/10.1016/j.jik.2023.100333. 2023/01/01/.
32. Karalis VD. The Integration of Artificial Intelligence into Clinical Practice. Appl Biosci. 2024;3(1):14–44. https://doi.org/10.3390/applbiosci3010002.
33. Podder I, Pipil N, Dhabal A, Mondal S, Pienyii V, Mondal H. Evaluation of Artificial Intelligence-based chatbot responses to common dermatological queries. Jordan Med J. 2024;07/20:58:271–7. https://doi.org/10.35516/jmj.v58i2.2960.
34. Javaid M, Haleem A, Singh RP. ChatGPT for healthcare services: An emerging stage for an innovative perspective. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*. 2023/02/01/ 2023;3(1):100105. https://doi.org/10.1016/j.tbench.2023.100105
35. Durmaz Engin C, Karatas E, Ozturk T. Exploring the role of ChatGPT-4, BingAI, and Gemini as virtual consultants to educate families about Retinopathy of Prematurity. Children. 2024;11(6):750. https://doi.org/10.3390/children11060750.
36. AlShehri Y, McConkey M, Lodhia P. ChatGPT has Educational potential: assessing ChatGPT responses to common patient hip arthroscopy questions. Arthroscopy. 2024. https://doi.org/10.1016/j.arthro.2024.06.017.
37. Roldan-Vasquez E, Mitri S, Bhasin S, et al. Reliability of artificial intelligence chatbot responses to frequently asked questions in breast surgical oncology. J Surg Oncol. 2024https://doi.org/10.1002/jso.27715.
38. Şahin B, Emre Genç Y, Doğan K, et al. Evaluating the performance of ChatGPT in Urology: a comparative study of Knowledge Interpretation and Patient Guidance. J Endourol. 2024. https://doi.org/10.1089/end.2023.0413.
39. Ghanem D, Shu H, Bergstein V, et al. Educating patients on osteoporosis and bone health: can ChatGPT provide high-quality content? Eur J Orthop Surg Traumatol. 2024https://doi.org/10.1007/s00590-024-03990-y.
40. Ding Q, Ding D, Wang Y, Guan C, Ding B. Unraveling the landscape of large language models: a systematic review and future perspectives. J Electron Bus Digit Econ. 2024;3(1):3–19. https://doi.org/10.1108/JEBDE-08-2023-0015.
41. Devlin J, Chang M-W, Lee K, Toutanova K, Bert. Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:181004805*. 2018;https://doi.org/10.48550/arXiv.1810.04805
42. Tanwar E, Borthakur M, Dutta S, Chakraborty T. Multilingual LLMs are Better Cross-lingual In-context Learners with Alignment. *arXiv preprint arXiv:230505940*. 2023;https://doi.org/10.48550/arXiv.2305.05940
43. Meskó B. The impact of Multimodal large Language models on Health Care's future. J Med Internet Res Nov. 2023;2:25:e52865. https://doi.org/10.2196/52865.
44. Dahal S, Aoun M. Exploring the Role of Machine Translation in Improving Health Information Access for Linguistically Diverse Populations. *Journal of Intelligent Information Systems*. 08/13. 2023;8:4–6. doi:n/a; https://questsquare.org/index.php/JOURNALAIIS/article/view/1
45. Ghaderzadeh M, Hosseini A, Asadi F, Abolghasemi H, Bashash D, Roshanpoor A. Automated detection model in classification of B-Lymphoblast cells from normal B-Lymphoid precursors in blood smear microscopic images based on the Majority Voting technique. Sci Program. 2022;4801671. https://doi.org/10.1155/2022/4801671. /01/04 2022;2022.

46. Malik YS, Sircar S, Bhat S, et al. How artificial intelligence may help the Covid-19 pandemic: pitfalls and lessons for the future. Rev Med Virol Sep. 2021;31(5):1–11. https://doi.org/10.1002/rmv.2205.
47. Madhu G, Mohamed AW, Kautish S, Shah MA, Ali I. Intelligent diagnostic model for malaria parasite detection and classification using imperative inception-based capsule neural networks. Sci Rep Aug. 2023;17(1):13377. https://doi.org/10.1038/s41598-023-40317-z.
48. Li J, Dada A, Kleesiek J, Egger J. ChatGPT in Healthcare: a taxonomy and systematic review. *medRxiv*. 2023:2023.03.30.23287899. https://doi.org/10.1101/2023.03.30.23287899
49. 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. /08/19 2023.
50. Wang Y, McKee M, Torbica A, Stuckler D. Systematic Literature Review on the spread of Health-related misinformation on Social Media. Soc Sci Med. 2019;240:112552. https://doi.org/10.1016/j.socscimed.2019.112552. 2019/11/01/.
51. Kleesiek J, Wu Y, Stiglic G, Egger J, Bian J. An opinion on ChatGPT in Health Care-written by humans only. J Nucl Med. May 2023;64(5):701–3. https://doi.org/10.2967/jnumed.123.265687.
52. Roumeliotis KI, Tselikas ND. ChatGPT and Open-AI models: a preliminary review. Future Internet. 2023;15(6):192. https://doi.org/10.3390/fi15060192.
53. Taye MM. Understanding of Machine Learning with Deep Learning: architectures, Workflow, applications and future directions. Computers. 2023;12(5):91. https://doi.org/10.3390/computers12050091.
54. Zhou W, He L, Nie X, et al. Accuracy and timeliness of knowledge dissemination on COVID-19 among people in rural and remote regions of China at the early stage of outbreak. Front Public Health. 2021;9:554038. https://doi.org/10.3389/fpubh.2021.554038.
55. Ramezani M, Takian A, Bakhtiari A, Rabiee HR, Ghazanfari S, Mostafavi H. The application of artificial intelligence in health policy: a scoping review. BMC Health Serv Res. 2023;2023/12/15(1):1416. https://doi.org/10.1186/s12913-023-10462-2.
56. Olawade DB, Wada OJ, David-Olawade AC, Kunonga E, Abaire O, Ling J. Using artificial intelligence to improve public health: a narrative review. Front Public Health. 2023;11:1196397. https://doi.org/10.3389/fpubh.2023.1196397.
57. Bautista YJP, Theran C, Aló R, Lima V. Health disparities through generative AI models: a comparison study using a Domain specific large Language Model. Springer Nature Switzerland; 2023. pp. 220–32.
58. Sallam M, Barakat M, Sallam M. METRICS: establishing a preliminary Checklist to standardize design and reporting of Artificial Intelligence-Based studies in Healthcare. JMIR Preprints. 2023. https://doi.org/10.2196/preprints.54704.
59. Sallam M, Barakat M, Sallam M. A preliminary Checklist (METRICS) to standardize the design and reporting of studies on generative Artificial Intelligence-based models in Health Care Education and Practice: Development Study Involving a Literature Review. Interact J Med Res Feb. 2024;15:13:e54704. https://doi.org/10.2196/54704.
60. Rosner B. Fundamentals of biostatistics. 8th ed. Cengage learning; 2015.
61. Centers for Disease Control and Prevention. Frequently Asked Influenza (Flu) Questions: 2022–2023 Season. 25 December 2023. 2023. 2023. https://www.cdc.gov/flu/season/faq-flu-season-2022-2023.htm
62. WHO Viet Nam. Q&A on COVID-19 and Breastfeeding. 25 December 2023. 2023. 2023. https://www.who.int/vietnam/news/feature-stories/detail/q-a-on-covid-19-and-breastfeeding
63. Centers for Disease Control and Prevention, Malaria. Frequently Asked Questions (FAQs). 25 December 2023, 2023. 2023. https://www.cdc.gov/malaria/about/faqs.html
64. Guinn KM, Rubin EJ. Tuberculosis: just the FAQs. mBio Dec. 2017;19(6). https://doi.org/10.1128/mBio.01910-17.
65. Rehman A, Ul-Ain Baloch N, Awais M. Practice of cupping (Hijama) and the risk of bloodborne infections. Am J Infect Control. 2014;42(10):1139. https://doi.org/10.1016/j.ajic.2014.06.031.
66. WHO South-East Asia. Post COVID-19 (long COVID) Q&A. 25 December 2023. 2023. 2023. https://www.who.int/southeastasia/outbreaks-and-emergencies/covid-19/questions/post-covid-19-q-a
67. The NHS website for England. Can you catch HIV from kissing? 25 December 2023, 2023. Updated 2021. https://www.nhs.uk/common-health-questions/sexual-health/can-you-catch-hiv-from-kissing/
68. The WHO Regional Office for the Eastern Mediterranean. Tuberculosis Frequently Asked Questions (FAQs). 25 December 2023. 2023. 2023. https://www.emro.who.int/tuberculosis/faqs/index.html
69. WHO. Coronavirus disease (COVID-19) and people living with HIV. 25 December 2023. 2023. Updated 7 June 2023. 2023. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/coronavirus-disease-(covid-19)-covid-19-and-people-living-with-hiv
70. Centers for Disease Control and Prevention. BCG Vaccine Fact Sheet. 25 December 2023. 2023. 2023. https://www.cdc.gov/tb/publications/factsheets/prevention/bcg.htm
71. Sallam M, Barakat M, Sallam M. Pilot testing of a Tool to standardize the Assessment of the Quality of Health Information Generated by Artificial Intelligence-based models. Cureus Nov. 2023;15(11):e49373. https://doi.org/10.7759/cureus.49373.
72. Chan PS-f, Fang Y, Cheung DH, et al. Effectiveness of chatbots in increasing uptake, intention, and attitudes related to any type of vaccination: a systematic review and meta-analysis. Appl Psychology: Health Well-Being. 2024/06/17 2024;n/a(n/a).
73. Haltaufderheide J, Ranisch R. The ethics of ChatGPT in medicine and healthcare: a systematic review on large Language models (LLMs). Npj Digit Med. 2024;07(1):183. https://doi.org/10.1038/s41746-024-01157-x. /08 2024.
74. Chen Y, Esmaeilzadeh P. Generative AI in Medical Practice: In-Depth exploration of privacy and Security challenges. J Med Internet Res Mar. 2024;8:26:e53008. https://doi.org/10.2196/53008.
75. Sallam M, Khalil R, Sallam M, Benchmarking Generative AI. A call for establishing a Comprehensive Framework and a generative AIQ test. Mesopotamian J Artif Intell Healthc. 2024;07/02:2024:69–75. https://doi.org/10.58496/MJAIH/2024/010.
76. Bala I, Pindoo I, Mijwil M, Abotaleb M, Yundong W. Ensuring security and privacy in Healthcare Systems: a Review Exploring challenges, solutions, Future trends, and the practical applications of Artificial Intelligence. Jordan Med J. 2024;07/15:2024. https://doi.org/10.35516/jmj.v58i2.2527.
77. Daneshjou R, Smith MP, Sun MD, Rotemberg V, Zou J. Lack of transparency and Potential Bias in Artificial Intelligence Data Sets and algorithms: a scoping review. JAMA Dermatol Nov. 2021;1(11):1362–9. https://doi.org/10.1001/jamadermatol.2021.3129.
78. Lai V, Ngo Trung N, Veyseh A, et al. ChatGPT Beyond English: towards a comprehensive evaluation of large Language models in Multilingual Learning. arXiv. 2023. https://doi.org/10.48550/arXiv.2304.05613.
79. Yeo YH, Samaan JS, Ng WH et al. GPT-4 outperforms ChatGPT in answering non-english questions related to cirrhosis. *medRxiv*. 2023:2023.05.04.23289482. https://doi.org/10.1101/2023.05.04.23289482
80. Fleisig E, Smith G, Bossi M, Rustagi I, Yin X, Klein D. Linguistic Bias in ChatGPT: Language models reinforce dialect discrimination. arXiv Preprint. 2024. https://doi.org/10.48550/arXiv.2406.08818.
81. Retzlaff N. Political biases of ChatGPT in different languages. Preprints: Preprints; 2024.
82. Liu X, Wu J, Shao A, et al. Uncovering Language disparity of ChatGPT on Retinal Vascular Disease Classification: cross-sectional study. J Med Internet Res Jan. 2024;22:26:e51926. https://doi.org/10.2196/51926.
83. Pugliese N, Polverini D, Lombardi R, et al. Evaluation of ChatGPT as a Counselling Tool for italian-speaking MASLD patients: Assessment of Accuracy, completeness and comprehensibility. J Personalized Med. 2024;14(6):568. https://doi.org/10.3390/jpm14060568.
84. Ghosh S, Caliskan A, Ignores Non-Gendered Pronouns. ChatGPT Perpetuates Gender Bias in Machine Translation and : Findings across Bengali and Five other Low-Resource Languages. presented at: Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society; 2023; Montr\'{e}al, QC, Canada. https://doi.org/10.1145/3600211.3604672
85. Taira K, Itaya T, Hanada A. Performance of the large Language Model ChatGPT on the National Nurse examinations in Japan: evaluation study. JMIR Nurs Jun 27. 2023;6:e47305. https://doi.org/10.2196/47305.
86. Watari T, Takagi S, Sakaguchi K, et al. Performance comparison of ChatGPT-4 and Japanese medical residents in the General Medicine In-Training examination: comparison study. JMIR Med Educ Dec. 2023;6:9:e52202. https://doi.org/10.2196/52202.
87. Vowels LM, Francois-Walcott RRR, Darwiche J. AI in relationship counselling: Evaluating ChatGPT's therapeutic capabilities in providing relationship advice. *Computers in Human Behavior: Artificial Humans*. 2024/08/01/ 2024;2(2):100078. https://doi.org/10.1016/j.chbah.2024.100078
88. Guigue P-A, Meyer R, Thivolle-Lioux G, Brezinov Y, Levin G. Performance of ChatGPT in French language Parcours d'Accès Spécifique Santé test and in OBGYN. Int J Gynecol Obstet. 2023https://doi.org/10.1002/ijgo.15083.

89. Morreel S, Mathysen D, Verhoeven V, Aye. AI! ChatGPT passes multiple-choice family medicine exam. *Medical Teacher*. 2023/06/03, 2023;45(6):665–666. doi: 10.1080/0142159X.2023.2187684.
90. Rosoł M, Gąsior JS, Łaba J, Korzeniewski K, Młyńczak M. Evaluation of the performance of GPT-3.5 and GPT-4 on the Polish Medical Final Examination. *Scientific Reports*. 2023/11/22 2023;13(1):20512. https://doi.org/10.1038/s41598-023-46995-z
91. Gobira M, Nakayama LF, Moreira R, Andrade E, Regatieri CVS, Belfort R. Jr. Performance of ChatGPT-4 in answering questions from the Brazilian National Examination for Medical Degree Revalidation. *Rev Assoc Med Bras (*1992*)*. 2023;69(10):e20230848. https://doi.org/10.1590/1806-9282.20230848
92. Samaan JS, Yeo YH, Ng WH et al. ChatGPT's ability to comprehend and answer cirrhosis related questions in Arabic. *Arab Journal of Gastroenterology*. 2023/08/01/, 2023;24(3):145–148. doi:10.1016/j.ajg.2023.08.001.
93. Banimelhem O, Amayreh W. Is ChatGPT a Good English to Arabic Machine Translation Tool? 2023:1–6.
94. Khondaker MTI, Waheed A, Nagoudi EMB, Abdul-Mageed M. GPTAraEval: a comprehensive evaluation of ChatGPT on Arabic NLP. arXiv Preprint arXiv:230514976. 2023. https://doi.org/10.48550/arXiv.2305.14976.
95. Yilmaz Muluk S, Olcucu N. The role of Artificial Intelligence in the primary Prevention of Common Musculoskeletal diseases. Cureus. 2024/7/25 2024;16(7):e65372. https://doi.org/10.7759/cureus.65372
96. Ferrara E. Should ChatGPT be biased? Challenges and risks of bias in large language models. First Monday. 2023;11/07(11). https://doi.org/10.5210/fm.v28i11.13346.
97. Gerke S, Minssen T, Cohen G. Chapter 12 - ethical and legal challenges of artificial intelligence-driven healthcare. In: Bohr A, Memarzadeh K, editors. Artificial Intelligence in Healthcare. Academic; 2020. pp. 295–336.
98. Khan B, Fatima H, Qureshi A et al. Drawbacks of Artificial Intelligence and their potential solutions in the Healthcare Sector. Biomed Mater Devices Feb 8 2023:1–8. https://doi.org/10.1007/s44174-023-00063-2
99. Crossnohere NL, Elsaid M, Paskett J, Bose-Brill S, Bridges JFP. Guidelines for Artificial Intelligence in Medicine: Literature Review and Content Analysis of frameworks. J Med Internet Res Aug. 2022;25(8):e36823. https://doi.org/10.2196/36823.
100. Sallam M, Mousa D. Evaluating ChatGPT performance in arabic dialects: a comparative study showing defects in responding to Jordanian and Tunisian general health prompts. Mesopotamian J Artif Intell Healthc. 2024;01/10:2024:1–7. https://doi.org/10.58496/MJAIH/2024/001.
101. Sousa VD, Rojjanasrirat W. Translation, adaptation and validation of instruments or scales for use in cross-cultural health care research: a clear and user-friendly guideline. J Eval Clin Pract Apr. 2011;17(2):268–74. https://doi.org/10.1111/j.1365-2753.2010.01434.x.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.