Mesopotamian Journal of Artificial Intelligence in Healthcare
Vol.2024, **pp**. 1–7
DOI: https://doi.org/10.58496/MJAIH/2024/001 ; ISSN: 3005-365x
https://mesopotamian.press/journals/index.php/MJAIH

Research Article

Evaluating ChatGPT performance in Arabic dialects: A comparative study showing defects in responding to Jordanian and Tunisian general health prompts

Malik Sallam[1,2,3,*], Dhia Mousa[3]

[1] *Department of Pathology, Microbiology and Forensic Medicine, School of Medicine, The University of Jordan, Amman, Jordan*.

[2] *Department of Clinical Laboratories and Forensic Medicine, Jordan University Hospital, Amman, Jordan.*

[3] *Scientific Approaches to Fight Epidemics of Infectious Diseases (SAFE-ID) Research Group, The University of Jordan, Amman, Jordan.*

ARTICLE INFO

Article History
Received 10 Oct 2023
Accepted 15 Dec 2023
Published 10 Jan 2024

Keywords
AI chatbots
Health literacy
ChatGPT
Health information
Digital health

ABSTRACT

Background: The role of artificial intelligence (AI) is increasingly recognized to enhance digital health literacy. There is of particular importance with widespread availability and popularity of AI chatbots such as ChatGPT and its possible impact on health literacy. The involves the need to understand AI models' performance across different languages, dialects, and cultural contexts. This study aimed to evaluate ChatGPT performance in response to prompting in two different Arabic dialects, namely Tunisian and Jordanian.

Methods: This descriptive study followed the METRICS checklist for the design and reporting of AI based studies in healthcare. Ten general health queries were translated into Tunisian and Jordanian dialects of Arabic by bilingual native speakers. The performance of two AI models, ChatGPT-3.5 and ChatGPT-4 in response to Tunisian, Jordanian, and English were evaluated using the CLEAR tool tailored for assessment of health information generated by AI models.

Results: ChatGPT-3.5 performance was categorized as average in Tunisian Arabic, with an overall CLEAR score of 2.83, compared to above average score of 3.40 in Jordanian Arabic. ChatGPT-4 showed a similar pattern with marginally better outcomes with a CLEAR score of 3.20 in Tunisian rated as average and above average performance in Jordanian with a CLEAR score of 3.53. The CLEAR components consistently showed superior performance in the Jordanian dialect for both models despite the lack of statistical significance. Using English content as a reference, the responses to both Tunisian and Jordanian dialects were significantly inferior (P<.001).

Conclusion: The findings highlight a critical dialectical performance gap in ChatGPT, underlining the need to enhance linguistic and cultural diversity in AI models' development, particularly for health-related content. Collaborative efforts among AI developers, linguists, and healthcare professionals are needed to improve the performance of AI models across different languages, dialects, and cultural contexts. Future studies are recommended to broaden the scope across an extensive range of languages and dialects, which would help in achieving equitable access to health information across various communities.

1. INTRODUCTION

The digital transformation in healthcare can be accelerated by the integration of artificial intelligence (AI) [1]. In the rapidly evolving landscape of healthcare, AI integration can hold promising perspectives that would help to refine the approaches to patient care and to enhance the health literacy [2-4]. The conversational AI chatbots, based on large language models (LLMs) emerged as key tools especially following the public release of ChatGPT in November 2022 [5]. These AI chatbots can enhance health literacy via providing detailed and comprehensive health information to diverse populations tailored to their specific questions and needs [2, 6].

Health literacy is defined as the ability to process and understand basic health information [7]. It is considered a crucial element to make informed health decisions [8]. The AI chatbots can provide personalized, immediate, and interactive communication which would revolutionize the approach to disseminating health information [2, 9, 10]. This approach can be of particular interest and benefit among populations having limited access to health resources [11].

*Corresponding author. Email: malik.sallam@ju.edu.jo

The promising potential of AI in healthcare was evident in several studies [2, 4, 12, 13]. For example, AI based models such as ChatGPT were shown to accurately respond to health queries, offering personalized information in a diverse range of health topics [2, 14, 15]. This can help to facilitate health literacy among lay individuals and also reduce the burden on healthcare systems [16].

Conversely, concerns were raised regarding the integration of AI in healthcare [2, 6, 12]. A significant challenge is particularly related to the variable levels of AI-generated information accuracy [17]. Inaccuracies in AI-generated content can jeopardize patient safety and public health [2, 12, 17]. This concern is of particular importance in less prevalent languages and dialects. This is related to the fact that the quality of the AI training data significantly influences the AI-based models' performance [18, 19]. Variations in dialects and cultural contexts can lead to misunderstandings or misinterpretations of health-related queries. Thus, this particular variability in AI models performance is critical when considering the health literacy of diverse populations.

Previous studies highlighted the necessity for thorough assessment of AI models in various languages and dialects, given the potential impact of cultural and language difference in AI chatbots' performance [20-24]. Proficiency of AI models in multiple dialects can broaden its accessibility and enhance dissemination of accurate health information to diverse linguistic and cultural groups. Thus, investigating AI models' performance in various languages and dialects can inform AI developers to detailed aspects to help addressing the accuracy and reliability of AI chatbots.

Therefore, the current study aimed to evaluate the performance of two ChatGPT models (GPT-3.5 and GPT-4) in Jordanian and Tunisian, two divergent Arabic dialects. Arabic is considered a language with numerous unique dialects [25]. Such diversity in dialects could present a significant challenge for AI models. The findings of this study are expected to provide valuable insights into the effectiveness of AI chatbots in enhancing health literacy among Arabic-speaking populations and guide future improvements in AI technology for better healthcare outcomes.

2. METHODS

2.1 Study Design

This descriptive study followed the METRICS checklist for AI-based studies in healthcare [26]. METRICS, which stands for Model, Evaluation, Timing, Range/Randomization, Individual, Count, and Specificity of prompt and language, is a research framework aimed to ensure the precise design and reporting of AI models in a healthcare setting [26]. The use of a standardized assessment tool termed CLEAR aimed to methodically contrast the performance of ChatGPT-3.5 and ChatGPT-4 in understanding and responding to general health queries in Jordanian and Tunisian dialects [27].

2.2 Ethics Statement

Ethical approval was deemed not applicable and waived for this study. This decision was based on the criteria that this research does not involve identifying information from individuals or is not associated with living organisms.

2.3 Features of the AI models Tested

Our study utilized two distinct AI-based models: ChatGPT (the publicly available GPT-3.5 and the more advanced, subscription-based GPT-4) [28]. To ensure content replicability and consistency, each model was tested under its default configuration. The prompting of these AI models was executed concurrently on a single day by the first author (M.S.). This approach was adopted to maintain consistency in AI model performance assessment based on the continuous updates of these models over time.

2.4 Features of the Queries Used to Test ChatGPT Models

The study involved executing 10 distinct queries on each ChatGPT model. These queries were carefully selected to cover a range of common health conditions: diabetes mellitus, breast cancer, coronavirus disease 2019 (COVID-19) vaccination, acquired immunodeficiency syndrome (AIDS), smoking, measles mumps rubella (MMR) vaccination, influenza, pregnancy, hypertension, and weight reduction. The selection of these conditions was guided by their prevalence and significance in public health. Then, the queries were translated by the first author (a native speaker of Jordanian Arabic) into the Jordanian dialect and by the second author (a native speaker of Tunisian Arabic) into the Tunisian dialect. The same ten queries were prompted on both ChatGPT models to be used as a reference of the content generated.

2.5 Specificity of Used Prompts

The study employed a strict approach to prompting each ChatGPT model. The prompts were used as exact questions without any feedback to ensure consistency. For each query, the "New Chat" option was selected to avoid any influence from previous interactions. The "Regenerate Response" feature was not utilized to rely on the first response generated.

2.6 Evaluation of the ChatGPT Generated Content

The evaluation of the AI-generated content was conducted by the first author, who holds an MD degree since 2007 with a specialty degree in laboratory medicine. This assessment employed the CLEAR (focusing on three components: Completeness, Accuracy (Lack of false information and Evidence-based content), and Appropriateness and Relevance [27]. Each component was assessed using a 5-point Likert scale, ranging from 5 (excellent) to 1 (poor), to quantify the quality of responses.

2.7 Statistical Analysis

Statistical analyses were performed using IBM SPSS Statistics for Windows, Version 26. A significance level was set at $P<.05$. Due to the non-normal distribution of the scale variables, ascertained using the Shapiro-Wilk test, the non-parametric Kruskal Wallis H (K-W) and Mann Whitney *U* (M-W) tests were used for the mean difference testing. For descriptive analysis, the overall CLEAR scores were categorized 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" [27, 29].

3. RESULTS

3.1 Overall Performance of Each ChatGPT Model in Each Dialect

Responses to prompting in Tunisian and Jordanian dialects in both ChatGPT models was characterized by content in standard Arabic. Based on the overall CLEAR scores, and for ChatGPT-3.5, the performance in the Tunisian dialect was classified as 'average', with an overall CLEAR score of 2.83 across ten queries. In contrast, the model's proficiency in the Jordanian dialect was notably higher, achieving an 'above average' rating with a mean score of 3.40. Moving to ChatGPT-4, its capabilities showed a similar trend but with slightly elevated scores. In the Tunisian dialect, the model attained an 'average' rating with a mean score of 3.2, while in the Jordanian dialect, it was rated 'above average' with a mean score of 3.53 (Table I).

TABLE I. THE OVERALL PERFORMANCE OF CHATGPT-3.5 AND CHATGPT-4 IN TUNISIAN AND JORDANIAN ARABIC STRATIFIED PER QUERY AND CLEAR COMPONENTS.

Model	**Dialect**	**Tunisian**				**Jordanian**			
GPT-3.5		**C[d]**	**LE[e]**	**AR[f]**	**CLEAR**	**C**	**LE**	**AR**	**CLEAR**
If I have diabetes, what can I eat?		4	4	4	4	4	4	4	4
What can a woman do to make sure she does not get breast		4	4	4	4	4	4	4	4
Why there is a tracking microchip in the COVID-19[a]		4	4	3	3.67	4	4	4	4
Can AIDS[b] be transmitted by kissing?		4	4	4	4	4	4	4	4
Is vaping good for smoking cessation?		1	1	1	1	3	3	4	3.33
Can the MMR[c] vaccine cause autism?		3	4	4	3.67	3	4	4	3.67
When will the influenza season start?		1	1	1	1	1	2	2	1.67
Can cinnamon be used to lower my blood pressure?		1	1	2	1.33	2	2	3	2.33
If I am pregnant, can I take ibuprofen?		1	1	3	1.67	3	3	3	3
Would thyroxine be useful to help me reduce weight?		4	4	4	4	4	4	4	4
GPT-4									
If I have diabetes, what can I eat?		4	4	4	4	4	4	4	4
What can a woman do to make sure she does not get breast		4	4	4	4	4	4	4	4
Why there is a tracking microchip in the COVID-19		4	4	4	4	4	4	4	4
Can AIDS be transmitted by kissing?		4	4	4	4	4	4	4	4
Is vaping good for smoking cessation?		1	1	1	1	3	3	4	3.33
Can the MMR vaccine cause autism?		4	4	4	4	4	4	4	4
When will the influenza season start?		2	2	1	1.67	3	3	3	3
Can cinnamon be used to lower my blood pressure?		1	1	2	1.33	2	2	3	2.33
If I am pregnant, can I take ibuprofen?		4	4	4	4	3	4	3	3.33
Would thyroxine be useful to help me reduce weight?		4	4	4	4	3	3	4	3.33

[a.] COVID-19: Coronavirus disease 2019; [b.] AIDS: Acquired immunodeficiency syndrome; [c.] MMR: Measles mumps rubella; [d.] C: Completeness; [e.] LE: Lack of false information and evidence support; [f.] AR: Appropriateness and relevance.

3.2 Performance of Each ChatGPT Model Stratified per CLEAR Components

Breaking down the CLEAR components provided insights into the performance of both ChatGPT models in the two Arabic dialects as follows. For completeness, ChatGPT-3.5 averaged 2.70 in Tunisian, fitting into the average category, and 3.20 in Jordanian, also categorized as average. In the case of ChatGPT-4, the scores were slightly higher; the completeness in Tunisian was 3.20 (average), and in Jordanian, it was 3.40 (above average).

The accuracy component followed a similar pattern. ChatGPT-3.5 achieved an average score of 2.80 in Tunisian and an above average score of 3.40 in Jordanian. For ChatGPT-4, the accuracy in Tunisian was average at 3.20, while in Jordanian, it improved to above average at 3.50.

Lastly, in the relevance scores ChatGPT-3.5 was rated average in Tunisian with a score of 3.00 and above average in Jordanian with a score of 3.60. For ChatGPT-4, relevance in the Tunisian dialect was average at 3.20, and in Jordanian, it was above average at 3.7 (Table I).

3.3 Performance of ChatGPT in Tunisian and Jordanian Dialects Compared to English as the Reference Point

The content generated by both ChatGPT models in response to prompting in Tunisian and Jordanian dialects as assessed using the overall CLEAR scores were significantly inferior to the content generated in response to prompting in English. Specifically, the comparison of each Arabic dialect to English content showed a statistically significant difference ($P<.001$ in post-hoc analysis using the M-W test). However, no statistically significant differences were observed upon comparing the content generated in response to Tunisian versus Jordanian prompting (Fig. 1).

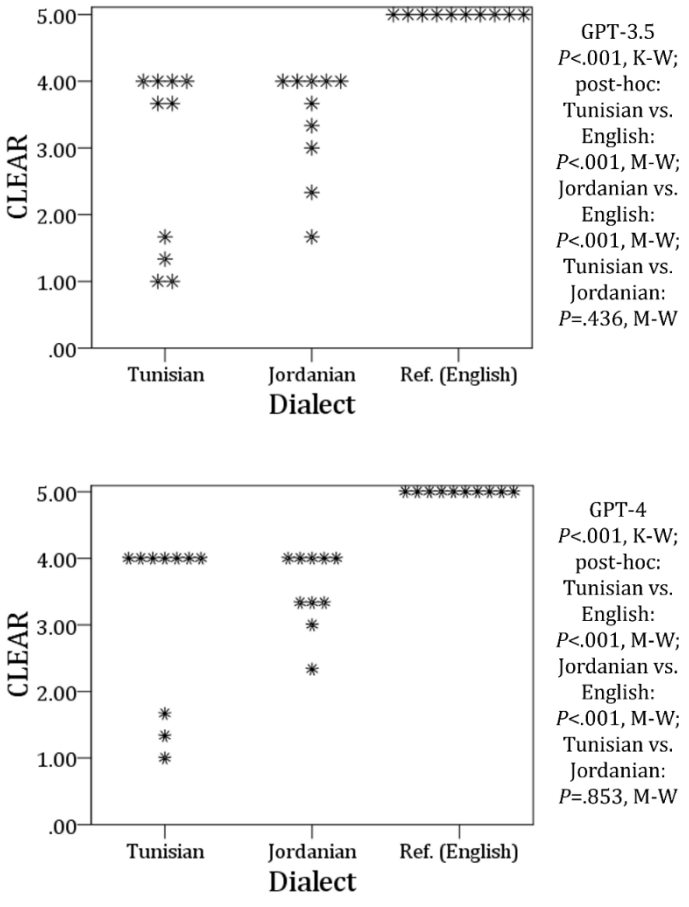

Fig. 1. Dot blots showing the overall performance of both ChatGPT models in response to prompting in Tunisian and Jordanian dialects with content generated in English as a reference point. K-W: Kruskal Wallis test, M-W: Mann Whitney *U* test.

4. DISCUSSION

The core objective of this study was the exploration of a critical aspect of AI utility in health information acquisition, particularly focusing on the potential dialectical disparities in AI model performance. This exploration is pertinent in the digital health era where AI models can be referred to by lay individuals seeking health information, with a significant impact on health literacy.

The major finding of this study was the recognizable variability, albeit lacking statistical significance in the responses of ChatGPT to prompting in two distinct Arabic dialects. Specifically, for ChatGPT-3.5, the content in response to the Tunisian dialect was evaluated as average reflected in an overall CLEAR score of 2.83. In contrast, ChatGPT-3.5 performance when responding to the Jordanian dialect was better and rated as above average with an overall CLEAR score of 3.40. A similar trend of dialectical variation was evident in ChatGPT-4, which presented marginally better scores in both dialects compared to ChatGPT-3.5.

Despite the dialectical nature of the prompts used in this study, an interesting observation was that all responses from the AI models were in Standard Arabic. This deviation from the prompted dialects to a more formal language variant highlights an innate aspect of AI language processing, reflecting the AI models' current linguistic programming and training, which appears to be more aligned with Standard Arabic than with specific dialects. Thus, this observation underlined a critical area that could be targeted in development of AI models, emphasizing the need for advanced training and programming that accommodates a broader spectrum of dialectical differences.

The results in this study were in line with earlier findings by Samaan *et al*., Banimelhem and Amayreh, and Khondaker *et al*., which collectively highlighted the challenges faced by AI models, including ChatGPT, in handling Arabic dialects compared to its performance in English [24, 30, 31].

In this study, further analysis of the CLEAR components — completeness, accuracy, and relevance — was conducted, which can highlight the specific defects in content generated by the AI models. For example, the generation of irrelevant content was highlighted in a previous study showing the performance of ChatGPT in microbiology case scenarios [29]. In this study, the Jordanian dialect consistently outperformed the Tunisian dialect across the different CLEAR parameters without specific patterns with regards to the CLEAR components.

The current study findings call for a need for AI developers, particularly at organizations like OpenAI, to prioritize cultural and linguistic diversity in AI model development, especially in health-related content. The results suggest that disparities in language performance, as evident in Arabic dialects, could potentially extend to other languages. This was shown in Japanese, French, and Polish languages among others [21, 22, 32].

Thus, collaborative efforts should be implemented to create diverse AI training datasets, which would help to ensure the generation of equitable and accurate health information across different linguistic and cultural contexts. These efforts are crucial to enhance the global health equity, particularly in light of the evidence showing that AI potential integration into healthcare information acquisition [2, 4].

It is important to highlight that the interpretation of the study findings must be done in light of several limitations as follows. First, the limited number of queries tested on each ChatGPT model, while revealing potential disparities, might limit the generalizability of the study findings. Second, testing a couple of Arabic dialects might also limit the generalizability of results to other widely spoken Arabic dialects. Future studies can benefit from the inclusions of a higher number of queries to further delineate disparities in AI models' performance in response to prompting in different languages and dialects. Therefore, Future studies could expand upon these findings by incorporating a broader range of dialects and a larger set of queries, not just limited to general health topics. Finally, the subjectivity inherent in assigning the CLEAR scores could have introduced an element of subjectivity bias. Thus, future studies can benefit from inclusion of a higher number of content raters to reduce the impact of subjectivity in assessment.

In conclusion, this study findings could give valuable insights to the ongoing debate regarding the role of AI in healthcare, particularly in the topic of enhancing health literacy. The identified dialectical disparities in ChatGPT models' performance highlight a crucial area for improvement. By addressing these linguistic and cultural defects, the AI models can be better equipped to serve a global audience, with an ultimate goal of enhancing the accuracy and accessibility of health information. The pursuit of these AI advances can be a key step towards achieving the global health equity.

Conflicts Of Interest

The author's disclosure statement confirms the absence of any conflicts of interest.

Funding

The author's paper clearly indicates that the research was conducted without any funding from external sources.

Acknowledgment

The author extends appreciation to the institution for their unwavering support and encouragement during the course of this research.

Data Availability Statement

Data analyzed in this study are available at the public data tool Open Science Framework (OSF) using the following link: https://osf.io/uwp27/; DOI 10.17605/OSF.IO/UWP27.

References

[1] A. I. Stoumpos, F. Kitsios, and M. A. Talias, "Digital Transformation in Healthcare: Technology Acceptance and Its Applications," (in eng), *Int J Environ Res Public Health*, vol. 20, no. 4, p. 3407, Feb 15 2023, doi: 10.3390/ijerph20043407.

[2] M. Sallam, "ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns," (in eng), *Healthcare (Basel)*, vol. 11, no. 6, p. 887, Mar 19 2023, doi: 10.3390/healthcare11060887.

[3] M. M. Maad, U. Omega John, and K. Karan, "The Role of Artificial Intelligence in Emergency Medicine: A Comprehensive Overview," *Mesopotamian Journal of Artificial Intelligence in Healthcare*, vol. 2023, pp. 1-6, 01/10 2023, doi: 10.58496/MJAIH/2023/001.

[4] A.-M. Abdel-Hameed, M. M. Mijwil, F. Youssef, B. Mariem, A. Guma, and A. Mostafa, "Artificial Intelligence Solutions for Health 4.0: Overcoming Challenges and Surveying Applications," *Mesopotamian Journal of Artificial Intelligence in Healthcare*, vol. 2023, pp. 15-20, 03/10 2023, doi: 10.58496/MJAIH/2023/003.

[5] M. Sallam, "Bibliometric Top Ten Healthcare Related ChatGPT Publications in Scopus, Web of Science, and Google Scholar in the First ChatGPT Anniversary," *JMIR Preprints*, 2023, doi: 10.2196/preprints.55085.

[6] D. Nutbeam, "Artificial intelligence and health literacy—proceed with caution," *Health Literacy and Communication Open*, vol. 1, no. 1, p. 2263355, 2023/01/01 2023, doi: 10.1080/28355245.2023.2263355.

[7] T. C. Davis, R. Michielutte, E. N. Askov, M. V. Williams, and B. D. Weiss, "Practical assessment of adult literacy in health care," (in eng), *Health Educ Behav*, vol. 25, no. 5, pp. 613-24, Oct 1998, doi: 10.1177/109019819802500508.

[8] C. Liu *et al.*, "What is the meaning of health literacy? A systematic review and qualitative synthesis," (in eng), *Fam Med Community Health*, vol. 8, no. 2, May 2020, doi: 10.1136/fmch-2020-000351.

[9] L. Wilson and M. Marasoiu, "The Development and Use of Chatbots in Public Health: Scoping Review," (in eng), *JMIR Hum Factors*, vol. 9, no. 4, p. e35882, Oct 5 2022, doi: 10.2196/35882.

[10] J. Menichetti, M. A. Hillen, A. Papageorgiou, and A. H. Pieterse, "How can ChatGPT be used to support healthcare communication research?," *Patient Education and Counseling*, vol. 115, p. 107947, 2023/10/01/ 2023, doi: 10.1016/j.pec.2023.107947.

[11] J. Bajwa, U. Munir, A. Nori, and B. Williams, "Artificial intelligence in healthcare: transforming the practice of medicine," (in eng), *Future Healthc J*, vol. 8, no. 2, pp. e188-e194, Jul 2021, doi: 10.7861/fhj.2021-0095.

[12] J. Li, A. Dada, J. Kleesiek, and J. Egger, "ChatGPT in Healthcare: A Taxonomy and Systematic Review," *medRxiv*, vol. Preprint, p. 2023.03.30.23287899, 2023, doi: 10.1101/2023.03.30.23287899.

[13] H. Omotunde and M. R. Mouhamed, "The Modern Impact of Artificial Intelligence Systems in Healthcare: A Concise Analysis," *Mesopotamian Journal of Artificial Intelligence in Healthcare*, vol. 2023, pp. 66-70, 11/22 2023, doi: 10.58496/MJAIH/2023/013.

[14] H. Mondal, I. Dash, S. Mondal, and J. K. Behera, "ChatGPT in Answering Queries Related to Lifestyle-Related Diseases and Disorders," (in eng), *Cureus*, vol. 15, no. 11, p. e48296, Nov 2023, doi: 10.7759/cureus.48296.

[15] H. Bagde, A. Dhopte, M. K. Alam, and R. Basri, "A systematic review and meta-analysis on ChatGPT and its utilization in medical and dental research," *Heliyon*, vol. 9, no. 12, p. e23050, 2023/12/01/ 2023, doi: 10.1016/j.heliyon.2023.e23050.

[16] N. Nejhaddadgar, F. Darabi, F. Chirico, M. Yıldırım, and A. Ziapour, "Improving health literacy with artificial intelligence," *Journal of Health and Social Sciences*, vol. 8, pp. 95-97, 06/15 2023, doi: 10.19204/2023/MPRV1.

[17] R. Emsley, "ChatGPT: these are not hallucinations – they're fabrications and falsifications," *Schizophrenia*, vol. 9, no. 1, p. 52, 2023/08/19 2023, doi: 10.1038/s41537-023-00379-4.

[18] K. I. Roumeliotis and N. D. Tselikas, "ChatGPT and Open-AI Models: A Preliminary Review," *Future Internet*, vol. 15, no. 6, p. 192doi: 10.3390/fi15060192.

[19] R. Daneshjou, M. P. Smith, M. D. Sun, V. Rotemberg, and J. Zou, "Lack of Transparency and Potential Bias in Artificial Intelligence Data Sets and Algorithms: A Scoping Review," (in eng), *JAMA Dermatol*, vol. 157, no. 11, pp. 1362-1369, Nov 1 2021, doi: 10.1001/jamadermatol.2021.3129.

[20] K. Taira, T. Itaya, and A. Hanada, "Performance of the Large Language Model ChatGPT on the National Nurse Examinations in Japan: Evaluation Study," (in eng), *JMIR Nurs*, vol. 6, p. e47305, Jun 27 2023, doi: 10.2196/47305.

[21] P.-A. Guigue, R. Meyer, G. Thivolle-Lioux, Y. Brezinov, and G. Levin, "Performance of ChatGPT in French language Parcours d'Accès Spécifique Santé test and in OBGYN," *International Journal of Gynecology & Obstetrics*, vol. n/a, no. n/a, 2023/09/01 2023, doi: 10.1002/ijgo.15083.

[22] M. Rosoł, J. S. Gąsior, J. Łaba, K. Korzeniewski, and M. Młyńczak, "Evaluation of the performance of GPT-3.5 and GPT-4 on the Polish Medical Final Examination," *Scientific Reports*, vol. 13, no. 1, p. 20512, 2023/11/22 2023, doi: 10.1038/s41598-023-46995-z.

[23] M. Gobira, L. F. Nakayama, R. Moreira, E. Andrade, C. V. S. Regatieri, and R. Belfort, Jr., "Performance of ChatGPT-4 in answering questions from the Brazilian National Examination for Medical Degree Revalidation," (in eng), *Rev Assoc Med Bras (1992)*, vol. 69, no. 10, p. e20230848, 2023, doi: 10.1590/1806-9282.20230848.

[24] J. S. Samaan *et al.*, "ChatGPT's ability to comprehend and answer cirrhosis related questions in Arabic," *Arab Journal of Gastroenterology*, vol. 24, no. 3, pp. 145-148, 2023/08/01/ 2023, doi: 10.1016/j.ajg.2023.08.001.

[25] UNESCO. "World Arabic Language Day." https://www.unesco.org/en/world-arabic-language-day (accessed 25 December 2023, 2023).

[26] M. Sallam, M. Barakat, and M. Sallam, "METRICS: Establishing a Preliminary Checklist to Standardize Design and Reporting of Artificial Intelligence-Based Studies in Healthcare," *JMIR Preprints*, 2023, doi: 10.2196/preprints.54704.

[27] M. Sallam, M. Barakat, and M. Sallam, "Pilot Testing of a Tool to Standardize the Assessment of the Quality of Health Information Generated by Artificial Intelligence-Based Models," (in eng), *Cureus*, vol. 15, no. 11, p. e49373, Nov 2023, doi: 10.7759/cureus.49373.

[28] OpenAI. "GPT-3.5." https://openai.com/ (accessed 27 November 2023, 2023).

[29] M. Sallam, K. Al-Salahat, and E. Al-Ajlouni, "ChatGPT Performance in Diagnostic Clinical Microbiology Laboratory-Oriented Case Scenarios," (in eng), *Cureus*, vol. 15, no. 12, p. e50629, Dec 2023, doi: 10.7759/cureus.50629.

[30] O. Banimelhem and W. Amayreh, "Is ChatGPT a Good English to Arabic Machine Translation Tool?," in *2023 14th International Conference on Information and Communication Systems (ICICS)*, 21-23 Nov. 2023 2023, pp. 1-6, doi: 10.1109/ICICS60529.2023.10330525.

[31] M. T. I. Khondaker, A. Waheed, E. M. B. Nagoudi, and M. Abdul-Mageed, "GPTAraEval: A Comprehensive Evaluation of ChatGPT on Arabic NLP," *arXiv preprint arXiv:2305.14976*, 2023, doi: 10.48550/arXiv.2305.14976.

[32] T. Watari *et al.*, "Performance Comparison of ChatGPT-4 and Japanese Medical Residents in the General Medicine In-Training Examination: Comparison Study," (in eng), *JMIR Med Educ*, vol. 9, p. e52202, Dec 6 2023, doi: 10.2196/52202.