The Performance of OpenAI ChatGPT-4 and Google Gemini in Virology Multiple-Choice Questions: A Comparative Analysis of English and Arabic Responses

Malik Sallam

malik.sallam@ju.edu.jo

The University of Jordan

Kholoud Al-Mahzoum

The University of Jordan

Rawan Ahmad Almutawaa

The University of Jordan

Jasmen Ahmad Alhashash

The University of Jordan

Retaj Abdullah Dashti

The University of Jordan

Danah Raed AlSafy

The University of Jordan

Reem Abdullah Almutairi

The University of Jordan

Muna Barakat

Applied Science Private University

Short Report

Keywords:

Posted Date: April 12th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-4220786/v1

Additional Declarations: No competing interests reported.

Abstract

Background: The integration of artificial intelligence (AI) in healthcare education is inevitable. Understanding the proficiency of generative AI in different languages to answer complex questions is crucial for educational purposes.

Objective: To compare the performance ChatGPT-4 and Gemini in answering Virology multiple-choice questions (MCQs) in English and Arabic, while assessing the quality of the generated content.

Methods: Both AI models' responses to 40 Virology MCQs were assessed for correctness and quality based on the CLEAR tool designed for evaluation of AI-generated content. The MCQs were classified into lower and higher cognitive categories based on the revised Bloom's taxonomy. The study design considered the METRICS checklist for the design and reporting of generative AI-based studies in healthcare.

Results: ChatGPT-4 and Gemini performed better in English compared to Arabic, with ChatGPT-4 consistently surpassing Gemini in correctness and CLEAR scores. ChatGPT-4 led Gemini with 80% vs. 62.5% correctness in English compared to 65% vs. 55% in Arabic. For both AI models, superior performance in lower cognitive domains was reported.

Conclusion: Both ChatGPT-4 and Gemini exhibited potential in educational applications; nevertheless, their performance varied across languages highlighting the importance of continued development to ensure the effective AI integration in healthcare education globally.

Introduction

Arabic is the native language for over 400 million people with a major role for effective communication in the majority of Middle East and North African countries [1]. Nevertheless, English is the official language of teaching and learning in healthcare education across a majority of Arab countries [2–4]. Therefore, university students are challenged by a linguistic shift at the start of healthcare education [5, 6].

The emergence of generative artificial intelligence (AI) models can provide support to bridge the linguistic challenge in healthcare for non-native English speakers [7–10]. However, the utility of AI models should be preceded by a critical evaluation of the reliability of its generated content especially in non-English languages, given the dominant training of large language models (LLMs) in English [11, 12]. If generative AI models are integrated in education, the bias toward English in LLM training could undermine efforts to achieve global educational equity [13, 14].

The integration of generative AI models into various aspects of daily life has been marked by a growing interest [15, 16]. This trend was particularly notable in healthcare, where AI models can increase the operational efficiency and improve the quality of delivery of patient care and healthcare education [7, 17, 18]. The potential benefits of AI in education are widely recognized; nevertheless, valid ethical concerns

are recognized besides the concern regarding inaccuracies reported for the AI-generated content [7, 19–21].

The use of multiple-choice questions (MCQs) in healthcare education is recognized as a reliable method to evaluate the students' achievement of learning outcomes [22, 23]. A relevant approach of classifying MCQs is the revised Bloom's taxonomy which is based on cognitive functions ranging from basic knowledge recall to the application of knowledge in problem-solving and systematic analysis of various concepts [24–26].

Since the utility and integration of generative AI in various aspects of healthcare education appears inevitable, it is important to consider the strengths and limitations of such innovative technology [18, 27]. This involves a thorough evaluation of the performance of the widely used AI chatbots, such as ChatGPT and Gemini, in different educational contexts [28–30]. Recent studies investigated the capability of different AI models to pass exams in different domains, with a wide variability in performance as reviewed recently by Newton and Xiromeriti [31]. This variability can be attributed to different factors, such as the AI model used, the prompting approach, and importantly the language(s) used in prompting [31, 32]. Such findings highlight the necessity for continued research to elucidate the determinants of AI models' performance, thereby informing the refinement of AI algorithms for improved performance and subsequent improved utility in various disciplines such as healthcare education [7, 33, 34].

Therefore, the current study aimed to compare the performance of two prominent AI models (ChatGPT-4 versus Gemini) in English and Arabic languages within the specialized field of Virology. The original hypothesis postulated that generative AI models' performance in English is superior to that in Arabic, inferred based on the presumed higher quality of training data available in English and based on the few reports describing this disparity in language performance [35, 36]. Highlighting the critical discrepancies in generative AI performance across languages can lead to identification of possible areas for improvement by AI developers.

Methods

Study design

The study utilized the METRICS checklist for the design and reporting of generative AI studies in healthcare [32]. The basis of the study was a randomly selected 40 Virology MCQs, used for testing of medical students during the period 2017–2022. The MCQs were fully designed by the first author (M.S.), with a PhD degree in Clinical Virology. The MCQs were original, without any copyright issues.

The MCQs were classified based on the revised Bloom's taxonomy into two cognitive levels: higher involving 20 "Remember" and "Understand" MCQs; and lower involving 20 "Apply" and "Analyze" MCQs classified based on a consensus between the first and senior authors. The MCQs were translated into Arabic by the first author and back translated into English by the senior author, both bilingual in English and Arabic.

The study was approved by the institutional review board (IRB) at the Faculty of Pharmacy – Applied Science Private University (reference number: 2024-PHA-5).

Models of generative AI tested, settings, and testing time

Two generative AI models were selected for testing based on their relevance, popularity, and advanced capabilities. The two models were ChatGPT-4 (OpenAI, San Francisco, CA) [29], and Gemini (Google, Mountain View, CA) [28].

We did not use the “regenerate response” or “modify response” features and refrained from providing any feedback for the two models to avoid feedback bias. Testing was conducted during 17 February–2 March 2024.

Prompt and language specificity

The following exact prompt was used: “For the following virology MCQ, please select the single most appropriate answer with an explanation for the rationale behind selecting this choice and excluding the other choices”. All MCQs were presented independently one-by-one in English and one-by-one in Arabic.

AI content evaluation approach and individual involvement in evaluation

First, we assessed the correctness of responses based on the key answers of the MCQs. Then, subjective evaluation of the AI generated content was based on a modified version of the CLEAR tool [37]. This involved assessing the content on three dimensions: (1) Completeness of the generated response; (2) Accuracy reflected by lack of false knowledge and the content being evidence-based; and (3) Appropriateness and relevance of content being easy to understand, well organized, and free from irrelevant content [37]. Each dimension was evaluated using a 5-point Likert scale, with 1 indicating “poor” and 5 representing “excellent” [37]. To enhance the objectivity of the assessment, a predefined list of criteria was established through discussions between the first and senior authors. Subsequently, the content produced by the two models underwent independent evaluation by the two raters. The CLEAR score for each piece of content was calculated by averaging the scores across the three assessed dimensions. The overall average CLEAR scores were then derived by averaging the scores assigned by the two raters.

Statistical and data analyses

The statistical analysis was conducted using IBM SPSS Statistics Version 26.0 (Armonk, NY: IBM Corp). To explore the associations between categorical variables, we employed two-sided Fisher’s exact test (FET), while the associations between the scale variable (CLEAR score) and categorical variables was assessed using the non-parametric the Mann–Whitney *U* test (M-W). The Kolmogorov-Smirnov test was employed to confirm the non-normality of the scale variable.

Results

General performance of ChatGPT-4 versus Gemini in English and Arabic

A higher percentage of correct responses for both generative AI models was observed in English compared to Arabic despite the lack of statistical significance. For Gemini in Arabic, the total number of correct responses was 22/40 (55.0%) while the correct responses for the same MCQs in English was 25/40 (62.5%, *P* = .650, FET). A similar trend was observed for ChatGPT-4 with correct responses in Arabic at 26/40 (65.0%) compared to 32/40 (80.0%) in English (*P* = .210, FET).

In Arabic, higher number of correct responses was seen in ChatGPT-4 compared to Gemini (65.0% vs. 55.0%, *P* = .494, FET) and similarly higher number of correct responses was observed in English for ChatGPT-4 compared to Gemini (80.0% vs. 62.5%, *P* = .137, FET).

Performance of ChatGPT-4 and Gemini based on revised Bloom's categories

In the evaluation of Gemini and ChatGPT-4 performance across the two revised Bloom's cognitive domains, the overall performance was consistently better in lower cognitive domain (Table 1).

Table 1
Performance of Gemini and ChatGPT-4 in English and Arabic in Virology multiple choice questions (MCQs) based on revised Bloom's cognitive categories.

Generative AI[1] model testing condition	Answer	Revised Bloom's category		*P* value[7]
		Lower cognitive[4]	Higher cognitive[6]	
		N[5] (%)	N (%)	
Gemini in Arabic	Correct	14 (70.0)	8 (40.0)	.111
	Incorrect	6 (30.0)	12 (60.0)	
Gemini in English	Correct	14 (70.0)	11 (55.0)	.514
	Incorrect	6 (30.0)	9 (45.0)	
ChatGPT-4 in Arabic	Correct	18 (90.0)	8 (40.0)	**.002**
	Incorrect	2 (10.0)	12 (60.0)	
ChatGPT-4 in English	Correct	17 (85.0)	15 (75.0)	.695
	Incorrect	3 (15.0)	5 (25.0)	
Overall[2] Gemini	Correct	28 (70.0)	19 (47.5)	.069
	Incorrect	12 (30.0)	21 (52.5)	
Overall ChatGPT-4	Correct	35 (87.5)	23 (57.5)	**.005**
	Incorrect	5 (12.5)	17 (42.5)	
Total[3] Arabic	Correct	32 (80.0)	16 (40.0)	**.001**
	Incorrect	8 (20.0)	24 (60.0)	
Total English	Correct	31 (77.5)	26 (65.0)	.323
	Incorrect	9 (22.5)	14 (35.0)	

[1]AI: Artificial intelligence; [2]Overall: The results in Arabic and English combined; [3]Total: The results of both AI models combined; [4]Lower cognitive: Includes the "Remember" and "Understand" categories of MCQs; [5]N: Number; [6]Higher cognitive: Includes the "Analyze" and "Apply" categories of MCQs; [7]*P* value: Calculated using the two-sided Fisher's exact test. Statistically significant *P* values are highlighted in bold style.

Within each revised Bloom's domain, ChatGPT-4 was marginally better in performance outscoring Gemini as follows: in lower cognitive domain, ChatGPT-4 had 35 correct responses (87.5%) vs. 28 Gemini correct responses (70.0%, $P = .099$, FET). In higher cognitive domain, ChatGPT-4 had 23 correct responses (57.5%) vs. 19 Gemini correct responses (47.5%, $P = .502$, FET).

Performance of ChatGPT-4 versus Gemini based on the average CLEAR scores

The performance of both generative AI models was superior in English as opposed to Arabic based on the average CLEAR scores as follows. For Gemini, the average CLEAR score was 3.48 ± 1.16 in Arabic compared to 4.00 ± 1.06 in English (*P* = .022, M-W, **Fig. 1A**). For ChatGPT-4, the average CLEAR score was 4.20 ± 0.74 in Arabic compared to 4.68 ± 0.57 in English (*P* = .001, M-W, **Fig. 1B**).

Figure 1. The average CLEAR scores for generative AI models' output regarding Virology multiple choice questions (MCQs). Gemini performance in Arabic vs. English (A); ChatGPT-4 performance in Arabic vs. English (B); Gemini vs. ChatGPT-4 performance in Arabic (C); Gemini vs. ChatGPT-4 performance in English (D).

Please insert Fig. 1 here

In Arabic, Gemini received lower average CLEAR scores compared to ChatGPT-4 (3.48 ± 1.16 vs. 4.20 ± 0.74, *P* = .005, M-W, **Fig. 1C**), and the same pattern was noticed for English language with Gemini having lower average CLEAR scores compared to ChatGPT-4 (4.00 ± 1.06 vs. 4.68 ± 0.57, *P* = .002, M-W, **Fig. 1D**).

Discussion

The current study focused on comparative analysis of the generative AI models Gemini and ChatGPT-4 abilities to answer Virology MCQs across English and Arabic languages. The findings revealed potential limitations inherent in the current versions of AI technologies which should be addressed prior to its incorporation in healthcare education especially for non-English speakers.

The results highlighted a discernible performance disparity between English and Arabic, with both AI models showing a lower accuracy in Arabic. This finding can be attributed to challenges encountered within LLMs' processing capabilities, particularly for languages that possess complex grammatical structures or languages with limited digital resources. The reduced accuracy in Arabic emphasizes the necessity for enriched training datasets that more comprehensively cover the linguistic diversity inherent in global languages. This comes in light of growing evidence of lower performance of different generative AI models in non-English languages.

In line with our observations, Samaan et al. reported ChatGPT lower accuracy in Arabic compared to English for cirrhosis-related queries [36]. Similarly, Banimelhem and Amayreh reported suboptimal English to Arabic translation capabilities for ChatGPT [38]. Additionally, a recent study showed the superior performance of four generative AI models in English compared to Arabic in infectious disease queries [39], while an earlier study showed the inferior performance of ChatGPT in general health queries in Arabic dialects [35]. Additionally, the inferior performance of AI chatbots was reported in other non-English languages including Chinese [40], Polish [41], and Spanish [42].

Interestingly, the study findings showed that both Gemini and ChatGPT-4 struggled with higher cognitive MCQs, which need advanced critical thinking and problem-solving skills. This limitation of generative AI performance is particularly relevant in healthcare education, where the ability to apply knowledge creatively and critically is essential [43]. The observed limitation raises concerns about the current reliability of AI as an educational tool, which was reported in the context of various AI chatbots [7, 18, 34, 44]. Collectively, these results highlight the critical areas for future development and improvement in AI training approaches.

Of note, this study highlighted ChatGPT-4 superior performance compared to Gemini in processing both Arabic and English languages across various cognitive levels, particularly emphasizing a pronounced advantage in addressing higher cognitive MCQs in Arabic. These findings might hint to OpenAI leading position in the development of LLMs, while also acknowledging the continued need for enhancements to improve performance in educational contexts.

Finally, this study showed the need for substantial improvements in generative AI training to enhance performance in non-English languages and in processing of higher-order cognitive queries. Addressing these challenges can improve the quality of AI-generated content and ensure its reliability, rendering AI chatbots as effective educational tools across diverse linguistic and cultural contexts.

The study limitations included the limited number of MCQs, which can restrict the scope of performance evaluation in this study. The subjective assessment of AI-generated content based on the CLEAR scores is another limitation highlighting the need for caution in interpretation. Additionally, this study focused solely on Virology MCQs, which may limit the generalizability of the findings to other healthcare disciplines. Moreover, the rapid evolution of LLMs highlights that the results may not fully reflect the evolving capabilities of the same generative AI models over time.

In conclusion, the study findings showed the capabilities and limitations of ChatGPT-4 and Gemini in the future of AI-assisted education. The variations in performance observed between languages and cognitive categories highlight the need for continued research, development, and optimization of generative AI models. A special attention should be paid into enhancing the linguistic diversity and cognitive understanding capabilities of generative AI models to achieve global educational equity.

Abbreviations

AI
Artificial intelligence
CLEAR
Completeness of content, Lack of false information in the content, Evidence supporting the content, Appropriateness of the content, and Relevance
FET
Two-sided Fisher’s exact test
LLMs

Large language models
MCQ
Multiple choice question
METRICS
Model, Evaluation, Timing, Range/Randomization, Individual factors, Count, and Specificity of prompts and language
M-W
Mann Whiteny *U* test

Declarations

Author contributions

Conceptualization: M.S.; Data curation: M.S., K.A.-M., Rawan Ahmad Almutawaa, J.A.A., R.A.D., D.R.A., Reem Abdullah Almutairi, M.B.; Formal analysis: M.S., K.A.-M., Rawan Ahmad Almutawaa, J.A.A., R.A.D., D.R.A., Reem Abdullah Almutairi, M.B.; Investigation: M.S., K.A.-M., Rawan Ahmad Almutawaa, J.A.A., R.A.D., D.R.A., Reem Abdullah Almutairi, M.B.; Methodology: M.S., K.A.-M., Rawan Ahmad Almutawaa, J.A.A., R.A.D., D.R.A., Reem Abdullah Almutairi, M.B.; Visualization: M.S.; Project administration: M.S.; Supervision: M.S.; Writing - original draft: M.S.; Writing - review & editing: M.S., K.A.-M., Rawan Ahmad Almutawaa, J.A.A., R.A.D., D.R.A., Reem Abdullah Almutairi, M.B.; All authors contributed to the article and approved the submitted version.

Funding

This research received no external funding.

Data availability

The corresponding author (M.S.) can provide the datasets used and analyzed for this study upon reasonable request.

Declarations

Ethics approval

The study was approved by the institutional review board (IRB) at the Faculty of Pharmacy – Applied Science Private University (reference number: 2024-PHA-5).

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

References

1. UNESCO. World Arabic Language Day 2023 [updated 18 December 2023; cited 2024 7 March 2024]. Available from: https://www.unesco.org/en/world-arabic-language-day.
2. Alhamami M, Almelhi A. English or Arabic in Healthcare Education: Perspectives of Healthcare Alumni, Students, and Instructors. J Multidiscip Healthc. 2021;14:2537-47. Epub 20210915. doi: 10.2147/jmdh.S330579.
3. Kaliyadan F, Thalamkandathil N, Parupalli SR, Amin TT, Balaha MH, Al Bu Ali WH. English language proficiency and academic performance: A study of a medical preparatory year program in Saudi Arabia. Avicenna J Med. 2015;5(4):140-4. doi: 10.4103/2231-0770.165126.
4. Alshareef M, Mobaireek O, Mohamud M, Alrajhi Z, Alhamdan A, Hamad B. Decision Makers' Perspectives on the Language of Instruction in Medicine in Saudi Arabia: A Qualitative Study. Health Professions Education. 2018;4(4):308-16. doi: 10.1016/j.hpe.2018.03.006.
5. Sabbour SM, Dewedar SA, Kandil SK. Language barriers in medical education and attitudes towards Arabization of medicine: student and staff perspectives. East Mediterr Health J. 2012;16(12):1263-71. Epub 20121204. doi: 10.26719/2010.16.12.1263.
6. Tayem Y, AlShammari A, Albalawi N, Shareef M. Language barriers to studying medicine in English: perceptions of final-year medical students at the Arabian Gulf University. East Mediterr Health J. 2020;26(2):233-8. Epub 20200224. doi: 10.26719/2020.26.2.233.
7. Sallam M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. Healthcare (Basel). 2023;11(6):887. Epub 20230319. doi: 10.3390/healthcare11060887.
8. Hwang SI, Lim JS, Lee RW, Matsui Y, Iguchi T, Hiraki T, et al. Is ChatGPT a "Fire of Prometheus" for Non-Native English-Speaking Researchers in Academic Writing? Korean J Radiol. 2023;24(10):952-9. doi: 10.3348/kjr.2023.0773.
9. Teixeira da Silva JA. Can ChatGPT rescue or assist with language barriers in healthcare communication? Patient Education and Counseling. 2023;115:107940. doi: 10.1016/j.pec.2023.107940.
10. Seetharaman R. Revolutionizing Medical Education: Can ChatGPT Boost Subjective Learning and Expression? J Med Syst. 2023;47(1):61. Epub 20230509. doi: 10.1007/s10916-023-01957-w.
11. Nicholas G, Bhatia A. Lost in Translation: Large Language Models in Non-English Content Analysis. arXiv preprint. 2023. doi: 10.48550/arXiv.2306.07377.
12. Lai VD, Ngo NT, Veyseh APB, Man H, Dernoncourt F, Bui T, et al. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. arXiv preprint. 2023. doi: 10.48550/arXiv.2304.05613.

13. Gurevich E, El Hassan B, El Morr C. Equity within AI systems: What can health leaders expect? Healthc Manage Forum. 2023;36(2):119-24. Epub 20221013. doi: 10.1177/08404704221125368.
14. Holstein K, Doroudi S. Equity and Artificial Intelligence in Education: Will" AIEd" Amplify or Alleviate Inequities in Education? arXiv preprint. 2021. doi: 10.48550/arXiv.2104.12920.
15. Chatterjee J, Dethlefs N. This new conversational AI model can be your friend, philosopher, and guide ... and even your worst enemy. Patterns (N Y). 2023;4(1):100676. Epub 20230113. doi: 10.1016/j.patter.2022.100676.
16. Kocoń J, Cichecki I, Kaszyca O, Kochanek M, Szydło D, Baran J, et al. ChatGPT: Jack of all trades, master of none. Information Fusion. 2023;99:101861. doi: 10.1016/j.inffus.2023.101861.
17. Alowais SA, Alghamdi SS, Alsuhebany N, Alqahtani T, Alshaya AI, Almohareb SN, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Medical Education. 2023;23(1):689. doi: 10.1186/s12909-023-04698-z.
18. Sallam M, Salim NA, Barakat M, Al-Tammemi AB. ChatGPT applications in medical, dental, pharmacy, and public health education: A descriptive study highlighting the advantages and limitations. Narra J. 2023;3(1):e103. doi: 10.52225/narra.v3i1.103.
19. Oniani D, Hilsman J, Peng Y, Poropatich RK, Pamplin JC, Legault GL, et al. Adopting and expanding ethical principles for generative artificial intelligence from military to healthcare. npj Digital Medicine. 2023;6(1):225. doi: 10.1038/s41746-023-00965-x.
20. Cappellani F, Card KR, Shields CL, Pulido JS, Haller JA. Reliability and accuracy of artificial intelligence ChatGPT in providing information on ophthalmic diseases and management to patients. Eye. 2024. doi: 10.1038/s41433-023-02906-0.
21. Emsley R. ChatGPT: these are not hallucinations – they’re fabrications and falsifications. Schizophrenia. 2023;9(1):52. doi: 10.1038/s41537-023-00379-4.
22. Kwon HJ, Chae SJ, Park JH. Educational implications of assessing learning outcomes with multiple choice questions and short essay questions. Korean J Med Educ. 2023;35(3):285-90. Epub 20230831. doi: 10.3946/kjme.2023.266.
23. Singh T. Principles of assessment in medical education: Jaypee Brothers Medical Publishers; 2021.
24. Stringer JK, Santen SA, Lee E, Rawls M, Bailey J, Richards A, et al. Examining Bloom’s Taxonomy in Multiple Choice Questions: Students’ Approach to Questions. Medical Science Educator. 2021;31(4):1311-7. doi: 10.1007/s40670-021-01305-y.
25. Bloom BS, Krathwohl DR. Taxonomy of Educational Objectives: The Classification of Educational Goals: Longmans, Green; 1956. 403 p.
26. Seaman M. BLOOM'S TAXONOMY: Its Evolution, Revision, and Use in the Field of Education. Curriculum and Teaching Dialogue. 2011;13(1/2):29-131A.
27. Reddy S. Generative AI in healthcare: an implementation science informed translational path on application, integration and governance. Implement Sci. 2024;19(1):27. Epub 20240315. doi: 10.1186/s13012-024-01357-9.

28. Google. Gemini 2024 [cited 2024 5 March 2024]. Available from: https://gemini.google.com/app.
29. OpenAI. GPT-4 2023 [cited 2024 5 March 2024]. Available from: https://openai.com/.
30. Rane N, Choudhary S, Rane J. Gemini versus ChatGPT: applications, performance, architecture, capabilities, and implementation. Journal of Applied Artificial Intelligence. 2024;5(1):69-93. doi: 10.48185/jaai.v5i1.1052.
31. Newton P, Xiromeriti M. ChatGPT performance on multiple choice question examinations in higher education. A pragmatic scoping review. Assessment & Evaluation in Higher Education.1-18. doi: 10.1080/02602938.2023.2299059.
32. 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. 2024;13:e54704. Epub 20240215. doi: 10.2196/54704.
33. Bandi A, Adapa PV, Kuchi YE. The Power of Generative AI: A Review of Requirements, Models, Input–Output Formats, Evaluation Metrics, and Challenges. Future Internet [Internet]. 2023; 15(8):[260 p.].
34. 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 Medical Journal. 2024;58(1). doi: 10.35516/jmj.v58i1.2285.
35. 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 Journal of Artificial Intelligence in Healthcare. 2024;2024:1-7. doi: 10.58496/MJAIH/2024/001.
36. Samaan JS, Yeo YH, Ng WH, Ting P-S, Trivedi H, Vipani A, et al. ChatGPT's ability to comprehend and answer cirrhosis related questions in Arabic. Arab Journal of Gastroenterology. 2023;24(3):145-8. doi: 10.1016/j.ajg.2023.08.001.
37. 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. 2023;15(11):e49373. Epub 20231124. doi: 10.7759/cureus.49373.
38. Banimelhem O, Amayreh W, editors. Is ChatGPT a Good English to Arabic Machine Translation Tool? 2023 14th International Conference on Information and Communication Systems (ICICS); 2023 21-23 Nov. 2023.
39. Sallam M, Al-Mahzoum K, Alshuaib O, Alhajri H, Alotaibi F, Alkhurainej D, et al. Superior Performance of Artificial Intelligence Models in English Compared to Arabic in Infectious Disease Queries. Research Square. 2024. doi: 10.21203/rs.3.rs-3830452/v1.
40. Liu X, Wu J, Shao A, Shen W, Ye P, Wang Y, et al. Uncovering Language Disparity of ChatGPT on Retinal Vascular Disease Classification: Cross-Sectional Study. J Med Internet Res. 2024;26:e51926. Epub 20240122. doi: 10.2196/51926.
41. 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;13(1):20512. doi: 10.1038/s41598-023-46995-z.

42. Guillen-Grima F, Guillen-Aguinaga S, Guillen-Aguinaga L, Alas-Brun R, Onambele L, Ortega W, et al. Evaluating the Efficacy of ChatGPT in Navigating the Spanish Medical Residency Entrance Examination (MIR): Promising Horizons for AI in Clinical Medicine. Clin Pract. 2023;13(6):1460-87. Epub 20231120. doi: 10.3390/clinpract13060130.
43. Jonathan MS, Andrew DO, Kamal RM, Iain C, Sandy O, Kevan C, et al. Critical thinking in healthcare and education. BMJ. 2017;357:j2234. doi: 10.1136/bmj.j2234.
44. Michel-Villarreal R, Vilalta-Perdomo E, Salinas-Navarro DE, Thierry-Aguilera R, Gerardou FS. Challenges and Opportunities of Generative AI for Higher Education as Explained by ChatGPT. Education Sciences. 2023;13(9):856. doi: 10.3390/educsci13090856.

Figures

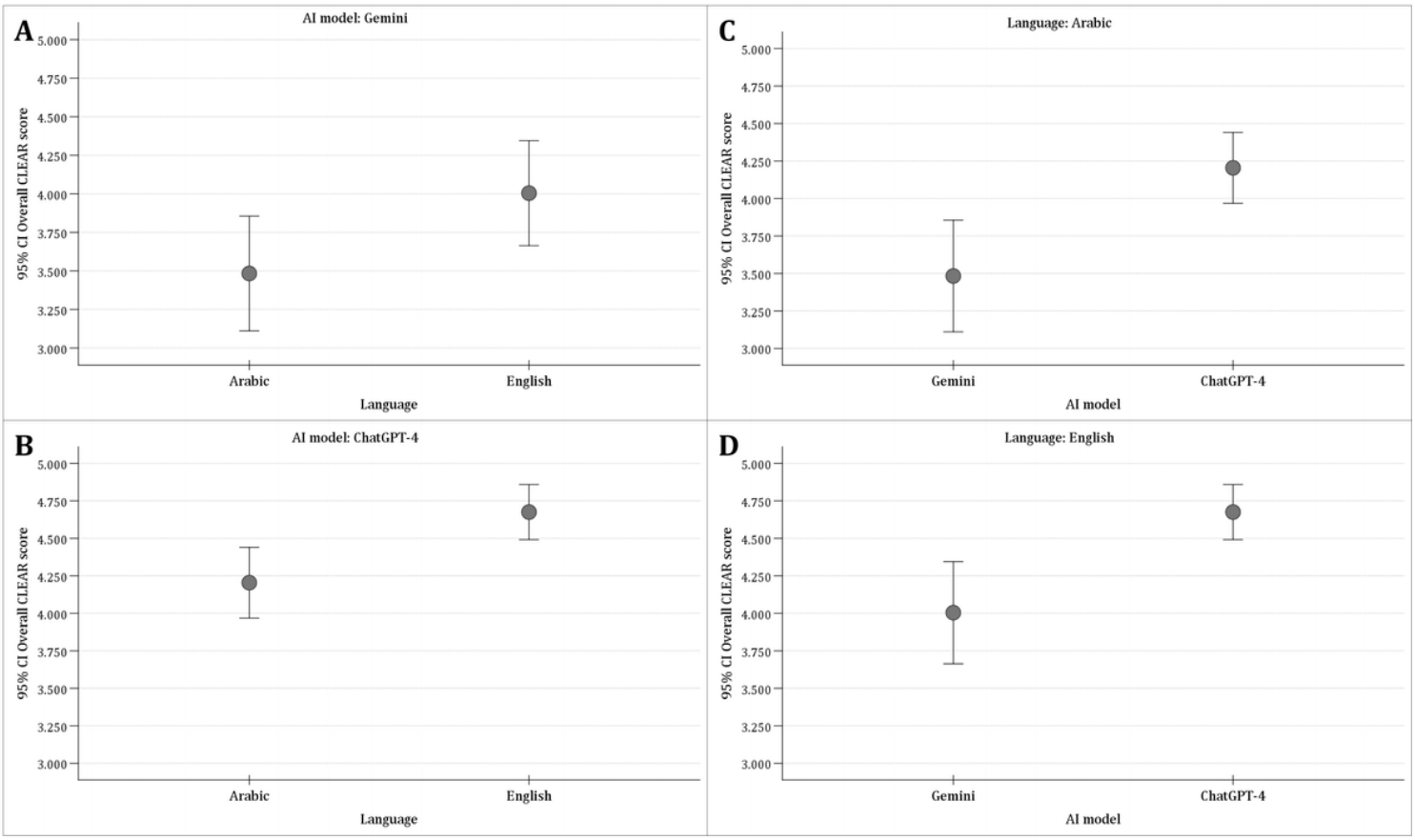

Figure 1

The average CLEAR scores for generative AI models' output regarding Virology multiple choice questions (MCQs). Gemini performance in Arabic vs. English (A); ChatGPT-4 performance in Arabic vs. English (B); Gemini vs. ChatGPT-4 performance in Arabic (C); Gemini vs. ChatGPT-4 performance in English (D).