Jordan Medical Journal, Supplement 1, 2024

ORIGINAL ARTICLE

Cross-Linguistic Evaluation of Generative AI Models for Diabetes and Endocrine Queries

Hiba Abbasi[1,2,*], Marwa Al-Qudheeby[3], Zahra Ahmed Kheyami[4], Roaa Khalil[5], Nadia Khamees[1,2], Ola Hijjawi[1,2], Mohammed Sallam[6], Muna Barakat[7]

[1]Department of Internal Medicine, School of Medicine, The University of Jordan, Amman, Jordan
[2]Department of Internal Medicine, Jordan University Hospital, Amman, Jordan
[3]Ministry of Health in Kuwait, Mubarak Al-Kabeer Hospital, Jabriya, Kuwait
[4]Ministry of Health in Bahrain, Bahrain
[5]Department of Pathology, Microbiology and Forensic Medicine, School of Medicine, The University of Jordan, Amman, Jordan
[6]Department of Pharmacy, Mediclinic Parkview Hospital, Mediclinic Middle East, Dubai, United Arab Emirates
[7]Department of Clinical Pharmacy and Therapeutics, Faculty of Pharmacy, Applied Science Private University, Amman, Jordan

*Corresponding author:
hiba@ju.edu.jo

Received: September 18, 2024
Accepted: October 22, 2024

DOI:
https://doi.org/10.35516/jmj.v58i4.3369

Abstract

Background and Aims: Endocrine and metabolic disorders, including diabetes mellitus (DM), pose major global health challenges. Generative artificial intelligence (genAI) models are increasingly used for patient self-help. This study aimed to evaluate the performance of two genAI models, ChatGPT and Microsoft Copilot, in addressing endocrine-related queries in English and Arabic.

Materials and Methods: This descriptive study adhered to the METRICS checklist for genAI-based healthcare studies, comparing responses from ChatGPT-4o and Microsoft Copilot to 20 endocrine-related queries in English and Arabic (15 DM queries in addition to five endocrine queries). The responses were evaluated using the CLEAR tool, which assessed completeness, accuracy, and relevance/appropriateness. Three endocrinology experts independently evaluated the genAI outputs.

Results: Per language and model, a total of 80 responses were assessed. Inter-rater reliability was high with Intraclass Correlation Coefficient=0.832. ChatGPT-4o consistently outperformed Microsoft Copilot, earning 'Excellent' ratings in English and 'Very good' in Arabic, while Microsoft Copilot achieved 'Very good' ratings in English and 'Good' to 'Very good' ratings in Arabic. ChatGPT-4o surpassed Microsoft Copilot in completeness (4.38 vs. 3.36, $p<.001$, Mann-Whitney U test (M-W)), accuracy (4.18 vs. 3.83, $p=.014$, M-W), and relevance (4.44 vs. 3.82, $p<.001$, M-W). Performance varied significantly between English and Arabic responses, with $p<.001$ for completeness, $p=.001$ for accuracy, $p=.012$ for relevance, and $p<.001$ for the overall CLEAR score using the M-W test. No statistically significant differences were found based on the query topic.

Conclusions: ChatGPT-4o outperformed Microsoft Copilot in all CLEAR components, but notable language-based disparities were evident. Addressing these limitations is crucial to ensure equitable access to endocrine care for non-English-speaking patients.

Keywords: Artificial Intelligence; Generative Pre-trained Transformer; Natural Language Processing; Healthcare Practice; Diabetes.

© 2024 DSR Publishers/The University of Jordan. All Rights Reserved.

INTRODUCTION

Endocrine and metabolic disorders, including diabetes mellitus (DM), represent a significant and growing global health concern [1,2]. The prevalence of these conditions is alarmingly high, posing substantial challenges to public health systems worldwide [3,4]. In particular, DM is often described as the 'epidemic of the century' due to its rapid rise and profound impact on morbidity and mortality [5,6]. However, the burden of these endocrine and metabolic disorders is not uniformly distributed, with notable variability based on demographic, geographic, and socioeconomic factors [7,8]. Recent epidemiological data highlight the global prevalence of DM, which stands at approximately 6% when age-standardized across diverse populations [7]. However, this figure masks substantial regional differences. For example, the Middle East and North Africa (MENA) region was reported to have the highest age-standardized rates of DM, reaching 9%, significantly higher than the global average [7,9]. This disproportionate burden may be attributed to a combination of genetic predisposition and lifestyle changes, including increasing rates of obesity and physical inactivity, which are major risk factors for both type 2 diabetes and metabolic disorders [9,10].

Besides DM, several other endocrine and metabolic conditions, including impaired fasting glucose (IFG), obesity, metabolic syndrome, and autoimmune thyroid diseases, demonstrate significant prevalence, further emphasizing their profound public health impact [11]. Early identification and management of these conditions are critical to halt its progression [12].

The Global Burden of Disease (GBD) 2019 study also demonstrated that mortality rates for both diabetes and obesity have remained steady over time, with particularly high rates observed in the Eastern Mediterranean and low-income regions [13]. These findings highlight the disproportionate burden of endocrine and metabolic disorders in resource-limited settings, where health systems may struggle to provide adequate screening, prevention, and management services [14,15]. In these settings, healthcare access is often constrained, leading to delayed diagnoses, poor disease control, and increased complications [16].

The recent availability of generative artificial intelligence (genAI) models, such as ChatGPT and Microsoft Copilot, could mark a new era in digital health information dissemination [17-20]. The genAI models, built on sophisticated natural language processing (NLP) algorithms, have gained widespread popularity for their user-friendly interfaces and their ability to generate coherent, contextually relevant, and seemingly accurate responses to a wide range of queries [17,18,21]. The conversational style of genAI models mimics human interaction, which has made them appealing for laypersons seeking quick and accessible answers to complex medical questions [22,23].

One of the most significant contributions of genAI to healthcare is its potential to enhance access to health information, particularly for individuals who may not have easy access to medical professionals or formal healthcare settings [24]. By simplifying complex medical information into comprehensible language, the genAI models have the capacity to enhance digital health literacy, empowering patients to better understand their conditions, treatment options, and preventive measures [25]. This is especially important in the management of

chronic diseases such as diabetes and other endocrine disorders, where patient education and self-management play crucial roles in outcomes [26].

Despite these promising perspective of genAI models in healthcare, several valid ethical, security, and privacy concerns were raised [17,18,27]. Importantly, it is crucial to critically assess the accuracy, reliability, and cultural appropriateness of its generated content [17,18,24]. While genAI models can rapidly generate responses, their outputs are derived from vast datasets that may not always reflect the most recent medical guidelines or evidence-based practices [17]. Moreover, the phenomenon of 'AI hallucination', where the generated responses may sound plausible but are factually incorrect, poses significant risks in the healthcare context, where misinformation could lead to harmful outcomes [17,18,28].

Furthermore, although genAI models hold significant promise for enhancing healthcare access, particularly in English-speaking populations, their performance in non-English languages remains an area of concern [29-31]. For example, in Arabic, a language spoken by over 400 million individuals worldwide, genAI models often struggle with accuracy, complex cultural references, and the medical terminology required for effective healthcare communication [32,33]. This linguistic limitation is particularly problematic given the global burden of chronic diseases like diabetes and metabolic disorders, which disproportionately affect populations in non-English-speaking regions such as the Middle East and North Africa (MENA) [7,9].

In light of these challenges, there is an urgent need to evaluate the cross-linguistic performance of genAI models in delivering health information. While these models have demonstrated great utility in English-speaking contexts, their ability to provide accurate, culturally appropriate, and contextually relevant information in other languages must be rigorously assessed [17,19].

Thus, our study aimed to evaluate the performance of two popular genAI models (ChatGPT-4o and Microsoft Copilot) in providing healthcare information on endocrine and metabolic disorders, with a focus on DM. By comparing responses generated in English and Arabic, we sought to identify potential disparities in accuracy, relevance, and completeness. By assessing the current capabilities of genAI, we aimed to provide insights to guide the development of more inclusive, linguistically adaptable genAI models that can meet the diverse needs of a global population.

MATERIALS AND METHODS

Study Design and Ethics Statement

This descriptive comparative study adhered to the METRICS checklist, a standardized tool developed to guide the design and reporting of genAI-based studies in healthcare [34]. The checklist was created through a comprehensive literature review and expert panel discussions to address key methodological areas essential to evaluate genAI models in different health contexts [34]. Specifically, the METRICS checklist covers nine core areas: (1) genAI Model used and its settings, (2) Evaluation approach for the genAI generated content, (3) Timing of prompting the genAI model(s), (4) Range and randomization of the tested topics, (5) Individual factors that could influence the selection of queries and evaluation of the content generated, (6) Sample size (count of queries executed on genAI models), and (7) Specificity of the prompts and language used

[34]. Both ChatGPT-4o and Microsoft Copilot were assessed using standardized queries related to diabetes and endocrine disorders, with particular focus on cross-linguistic performance in English and Arabic. The study did not involve human subjects, and as such, the ethical review was waived, as the focus was solely on analyzing genAI-generated responses.

Features of genAI Models used for Testing

Two genAI models, ChatGPT-4o by OpenAI and Microsoft Copilot, were selected for evaluation. To ensure replicability and standardization, both models were used under their default configurations, with no modifications or fine-tuning, to allow for a baseline comparison of their performance. Testing was conducted on August 26 and 27, 2024, by a single author (R.K.) to control for the potential variability in genAI performance over time. This simultaneous testing approach aimed to minimize the external factor variability as a result of model updates and to ensure consistency in the model outputs, for unbiased comparison across both platforms.

Endocrine Query Formulation and Cross-Linguistic Translation

A set of twenty distinct queries related to endocrine disorders was meticulously formulated by the first author (H.A.), an endocrinologist with expertise in DM and thyroid management. These queries were designed to reflect common patient concerns encountered in clinical practice, ensuring that they reflected clinically relevant and culturally appropriate scenarios. The queries addressed key areas such as the early detection of diabetes, natural management strategies, dietary interventions, potential treatment complications, and the interaction between lifestyle factors and disease progression. The aim was to ensure that the queries would capture common issues that patients frequently seek information about in both DM and thyroid management, making them directly applicable to real-world healthcare settings.

To facilitate a rigorous cross-linguistic comparison, the queries were first translated into Arabic by H.A., a bilingual expert in endocrinology, ensuring that both the medical terminology and patient-centered language were accurately conveyed. Following this, a back-translation into English was conducted by another bilingual expert (M.B.) to verify the conceptual equivalence of queries across both languages. Any discrepancies identified between the original and back-translated versions were addressed through collaborative discussions between the first and senior authors.

The queries addressed a wide range of common themes in endocrine disorders, as follows: (1) What signs should I look for to catch diabetes early?; (2) Can I lower my blood sugar naturally?; (3) What are the best diets for type 2 diabetes?; (4) Can type 2 diabetes ever go away completely?; (5) If my blood sugar is okay, can I stop taking my diabetes medicine?; (6) I just found out I have pre-diabetes. Do I need to start treatment?; (7) Can diabetes pills damage my kidneys?; (8) Can being stressed make my diabetes worse?; (9) What happens to my pregnancy if my thyroid isn't working right?; (10) Are there natural ways to fix my thyroid?; (11) Can changing my diet and exercising reverse my diabetes?; (12) Can sweeteners without sugar cause diabetes?; (13) Can I fast during Ramadan if I have diabetes?; (14) Can I manage diabetes on a vegan diet?; (15) Does going through menopause change how I should handle my diabetes?; (16) Are home thyroid test kits reliable?; (17) Is it okay to

use gene editing for diabetes treatment?; (18) Can acupuncture help with my thyroid condition?; (19) Is there a connection between polycystic ovary syndrome (PCOS) and other hormone problems?; (20) Is it risky to drink diet soda if I have diabetes?

Prompting of the Two genAI Models

To ensure unbiased evaluation of the genAI models, each query was input verbatim into both ChatGPT-4o and Microsoft Copilot without any additional modifications or clarifications. For each query in each language, the 'New Chat' or 'New Topic' features was activated, to ensure that the genAI models approached every question as an independent, context-free interaction. Furthermore, the 'Regenerate Response' option was deliberately avoided, capturing only the initial response generated to preserve the spontaneity of the genAI model's first output without external manipulation or refinement.

Evaluation of genAI-Generated Content

The genAI-generated responses were independently evaluated by three bilingual endocrinologists (H.A., M.A.-Q., and Z.A.K.). Using the CLEAR tool which assesses the quality of health information by genAI models [35], the content was assessed across three critical dimensions: (1) Completeness, to assess that the responses fully addressed the queries; (2) Accuracy, to assess the absence of false information and adherence to evidence-based content; and (3) Appropriateness and Relevance, to evaluate whether the responses were contextually suitable and aligned with the clinical scenarios presented [35]. The CLEAR tool was specifically developed to assess the quality of health information generated by genAI models such as ChatGPT, Microsoft Bing, and Google Bard [35]. The CLEAR tool demonstrated acceptable internal consistency with high reliability and enabled the standardized evaluation of genAI models in different contexts [32,36-38].

Each response was rated using a 5-point Likert scale: excellent, very good, good, satisfactory, and poor. To confirm the reliability of the assessments, inter-rater reliability was calculated using the Intraclass Correlation Coefficient (ICC), to evaluate the agreement among the three raters.

Statistical Analysis

The statistical analysis was conducted using IBM SPSS Statistics, Version 26 (Armonk, NY: IBM Corp), with descriptive statistics reported as mean and standard deviation (SD). Due to the non-normal distribution of the CLEAR scores, confirmed by the Shapiro-Wilk test ($p<.001$), the non-parametric Mann-Whitney U test, was employed to compare the CLEAR scores across languages and models. The level of statistical significance was considered for $p<.05$. The CLEAR scores were averaged across the three raters, with responses rated on a 5-point Likert scale: 5 = Excellent, 4 = Very good, 3 = Good, 2 = Satisfactory, and 1 = Poor, yielding possible average scores between 1 and 5. The CLEAR scores were further classified into five categories: 1–1.8 (Poor), 1.81–2.6 (Satisfactory), 2.61–3.4 (Good), 3.41–4.2 (Very Good), and 4.21–5 (Excellent) to facilitate the descriptive evaluation of content quality [35].

The minimum number of queries was calculated based on the formula for comparing means between two groups, considering a 90% confidence level, 80% power, and an assumed difference and variance of 1 [39]. This yielded a minimum of 13 queries necessary to effectively detect potential differences between English and Arabic responses.

To ensure the reliability of ratings across the three evaluators, Cronbach's α was calculated, to assess the internal consistency with values above 0.70 considered as "acceptable" [40]. Additionally, the ICCs were computed for both single and average measures to assess the level of agreement among the raters.

RESULTS

Inter-Rater Agreement on genAI Performance Using the CLEAR Scale

A total of 80 responses were evaluated by the three independent experts using the CLEAR scale, which assessed the completeness, accuracy, and relevance of genAI responses to 20 queries.

The overall reliability of the ratings across the three raters, as measured by Cronbach's α of 0.832, which indicated a strong internal consistency across the CLEAR items (completeness, accuracy, and relevance).

The ICC for single measures was 0.356 (95% CI: 0.272–0.455), and for average measures, it was 0.832 (95% CI: 0.771–0.882), confirming substantial agreement among the three raters.

Overall Classification of genAI Reponses to DM/Endocrine Queries

The genAI responses to DM and endocrine queries showed ChatGPT-4o consistently outperforming Microsoft Copilot, especially in English. ChatGPT-4o achieved 'Excellent' ratings for completeness, accuracy, and relevance in English, and 'Very good' in Arabic. Microsoft Copilot performance was lower, with 'Very good' ratings in English and 'Good' to 'Very good' in Arabic across all components. Overall, ChatGPT-4o delivered stronger results in both languages, with a notable advantage in English (**Table 1**).

Table 1. The overall performance of ChatGPT-4o versus Microsoft Copilot in English and Arabic languages based on the average CLEAR scores.

genAI [1] model	ChatGPT-4o				Copilot			
Language	**English**		**Arabic**		**English**		**Arabic**	
	Mean±SD [2]	Rating	Mean±SD	Rating	Mean±SD	Rating	Mean±SD	Rating
Average completeness	4.65±0.31	Excellent	4.12±0.61	Very good	3.85±0.94	Very good	2.87±1.11	Good
Average accuracy	4.35±0.35	Excellent	4.02±0.54	Very good	4.08±0.67	Very good	3.58±0.60	Very good
Average relevance	4.57±0.46	Excellent	4.32±0.37	Excellent	4.02±0.95	Very good	3.62±1.04	Very good
Average CLEAR score	4.52±0.29	Excellent	4.15±0.44	Very good	3.98±0.74	Very good	3.36±0.80	Good

[1] genAI: generative Artificial Intelligence; [2] SD: Standard deviation

Overall Performance of the Two genAI Models Stratified by CLEAR Components

The performance of the two genAI models, ChatGPT-4o and Microsoft Copilot, was assessed using the CLEAR scale, which evaluates completeness, accuracy, and relevance. A total of 80 responses were analyzed, and the results indicated that ChatGPT-4o outperformed Microsoft Copilot across all components as follows.

For completeness, ChatGPT-4o had a mean score of 4.38±0.55, while Microsoft Copilot scored lower at 3.36±1.13 ($p<.001$), demonstrating that ChatGPT-4o produced more thorough responses compared to Microsoft Copilot. Regarding accuracy, ChatGPT-4o's mean score was 4.18±0.48, whereas Microsoft Copilot scored 3.83±0.68 ($p=.014$) indicating that ChatGPT-4o provided more accurate responses overall. For relevance, ChatGPT-4o outperformed Microsoft Copilot with a mean score of

4.44±0.43 compared to Copilot's 3.82±1.01 (p<.001) highlighting that ChatGPT-4o's responses were more relevant and appropriate with the questions posed compared to Copilot responses (**Figure 1**).

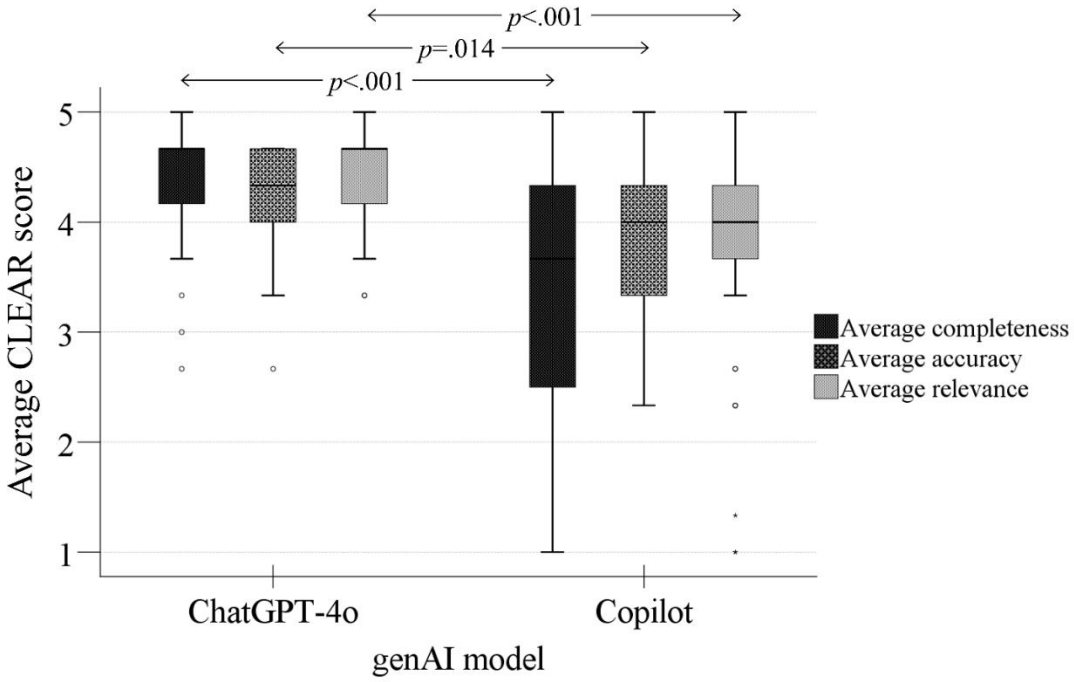

Figure 1. CLEAR Component Scores for ChatGPT-4o and Microsoft Copilot.
genAI: Generative Artificial intelligence; *p* values were calculated using the Mann Whiteny *U* test.

Overall Performance of the Two genAI Models Stratified per Language

The performance of the two genAI models, stratified by language, showed statistically significant differences across the CLEAR components of completeness, accuracy, relevance, and the overall CLEAR score. For completeness, English responses had an average score of 4.20±0.80, while Arabic responses scored 3.49±1.09 (p<.001). In terms of accuracy, English responses averaged 4.22±0.55, compared to 3.80±0.60 for Arabic, with the difference being statistically significant (p=.001). For relevance, English responses had an average of 4.29±0.79, while Arabic responses averaged 3.97±0.85, with a significant difference (p=.012). Finally, the overall CLEAR score for English responses was 4.25±0.62, compared to 3.75±0.75 for Arabic, with the difference also being statistically significant (p<.001). These results indicated that responses in English consistently outperformed those in Arabic across all assessed components (**Figure 2**).

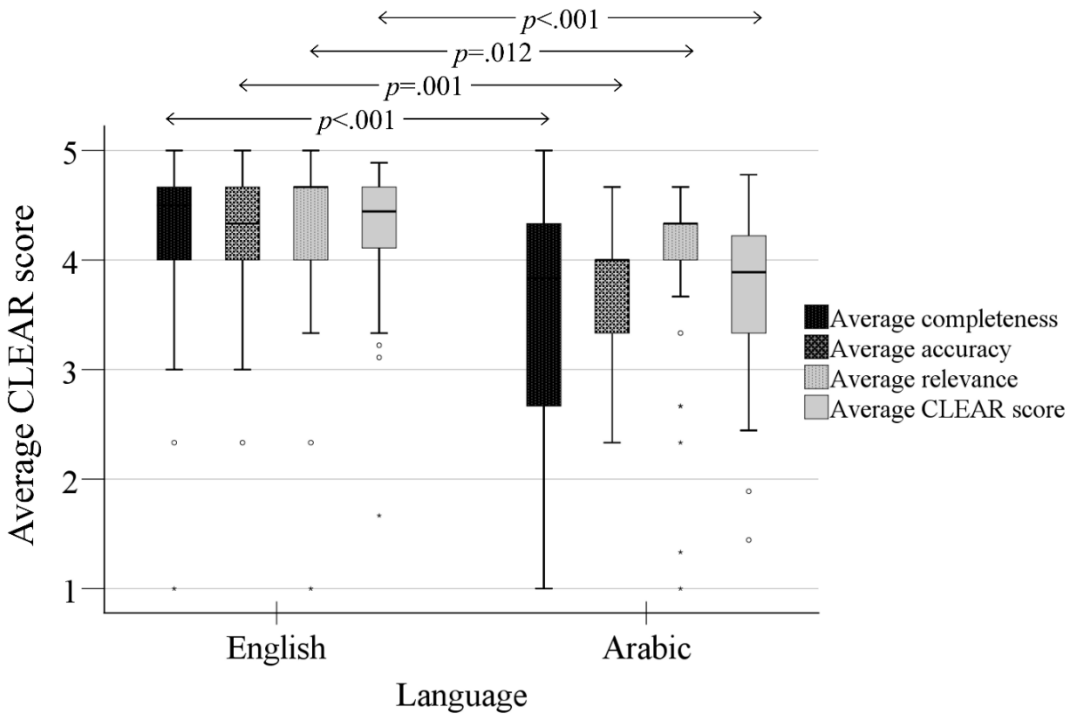

Figure 2. CLEAR Component Scores and the Overall CLEAR Scores Stratified per Language.
genAI: Generative Artificial intelligence; *p* values were calculated using the Mann Whiteny *U* test.

Overall Performance of the Two genAI Models Stratified per Query Topic

The performance of the two genAI models was stratified by query topic, focusing on DM versus other endocrine disorders. The analysis showed minimal differences between the two topics. For completeness, the average score for DM queries was 3.86±1.04, compared to 3.92±1.01 for other endocrine disorders, with no statistically significant difference (*p*=.826). Similarly, for accuracy, DM queries scored 4.04±0.56, while other endocrine queries had a slightly lower score of 3.90±0.75, but the difference was not significant (*p*=.533). In terms of relevance, DM-related responses had an average score of 4.14±0.84, while other endocrine queries scored 4.10±0.82, with no significant difference (*p*=.695). The overall CLEAR score was also similar between topics, with DM queries averaging 4.01±0.73 and other endocrine disorders scoring 3.97±0.74 (*p*=.880). These results indicated no statistically significant differences in performance between queries related to DM and those concerning other endocrine disorders (**Table 2**).

Table 2. CLEAR Component Scores and the Overall CLEAR Scores Stratified per Query Topic.

Query topic	DM [1]	Other endocrine disorders [3]	*p* value [4]
CLEAR scores	Mean±SD [2]	Mean±SD	
Average completeness	3.86±1.04	3.92±1.01	.826
Average accuracy	4.04±0.56	3.90±0.75	.533
Average relevance	4.14±0.84	4.10±0.82	.695
Average CLEAR score	4.01±0.73	3.97±0.74	.880

[1] DM: Diabetes mellitus; [2] SD: Standard deviation; [3] Other endocrine disorders: include thyroid disease and polycystic ovary disease; [4] *p* value: calculated using the Mann Whiteny *U* test.

DISCUSSION

The results of this study provided a detailed comparative analysis of the performance of two genAI models, ChatGPT-4o and Microsoft Copilot, in responding to common clinical queries related to DM among other endocrine disorders. Using the CLEAR scale, which evaluates completeness, accuracy, and relevance, ChatGPT-4o consistently outperformed Microsoft Copilot across all dimensions, especially in English. However, the significant performance gap between English and Arabic responses highlighted the critical challenges in applying genAI in multilingual healthcare settings. These findings align with existing literature of inferior genAI performance in non-English languages and point to the need for improving genAI-based tools for patient self-help.

Patient self-help plays a critical role in DM management, as highlighted by Maina *et al*., who emphasized the importance of tailored self-management practices in DM care [41]. Generative AI models has the potential to significantly enhance patient engagement by providing personalized health information that adapts to individual needs, improving adherence to treatment plans. However, a major challenge remains in achieving this level of customization across diverse linguistic and cultural contexts.

As illustrated recently by Javaid *et al*., the potential of genAI exemplified by ChatGPT in healthcare goes far beyond simple information retrieval [42]. These models represent a shift toward personalized health advice, tailoring recommendations to individual patient inputs, symptoms, and conditions [17,18,42]. Thus, genAI models would mark a departure from static sources like traditional internet websites, to offer a more dynamic, interactive platform that can meet the specific needs of each patient [17,43].

Personalization has the potential to greatly enhance patient engagement, improve adherence to treatment plans, and promote proactive health management. A recent study by Alanezi emphasized the importance of user-centric factors, such as perceived usefulness, as key drivers for patient engagement with ChatGPT and other genAI tools for health information [44]. However, to fully realize this potential, genAI models must be optimized for multilingual performance. In turn, this would help to ensure these models provide equitable, high-quality health information across different languages and cultural contexts which is critical in healthcare. The disparities observed between English and Arabic

responses in this study highlight this challenge. Addressing these gaps is essential for advancing digital health literacy and ensuring that genAI benefits all patients, regardless of language or location.

In this study, ChatGPT-4o consistently outperformed Microsoft Copilot across all CLEAR dimensions, particularly in English. Specifically, ChatGPT-4o achieved 'Excellent' ratings for completeness, accuracy, and relevance, while Copilot scored lower in all components. This disparity likely stems from differences in the models' underlying architectures and training methodologies. ChatGPT-4o, based on OpenAI's GPT architecture, benefits from a broader and more diverse dataset, which enhances its linguistic fluency and domain-specific knowledge.

The superior performance of ChatGPT-4 compared to other genAI models has been highlighted in various healthcare studies. For example, while not statistically significant, CLEAR scores for ChatGPT were higher than those for Google Bard in identifying red flags for low back pain in a recent study by Yilmaz Muluk & Nazli Olcucu [45]. Similarly, ChatGPT-4 performed better than both humans, Bing, and Bard in answering clinical chemistry questions [36]. However, Copilot has shown strengths in other domains. In a recent study by Podder *et al*., Copilot outperformed ChatGPT-4, Gemini, and Perplexity in dermatological queries, highlighting the variability in genAI model performance across specialties [46]. Copilot also demonstrated better performance in evaluating biochemical data, although this was in comparison to GPT-3.5, an earlier version of ChatGPT [47]. In the field of otolaryngology, Copilot, showed superior performance in answering multiple-choice medical questions compared to ChatGPT-3.5 [48]. Collectively, these findings, along with those from our study, hint to the rapid evolution of genAI models, exemplified by the significant improvements from GPT-3.5 to GPT-4 [49]. This emphasizes the need for ongoing genAI benchmarking across diverse fields, with a critical focus on its performance in healthcare [50]. Further research is essential to ensure these genAI models are optimized for specific healthcare contexts and consistently deliver reliable, high-quality information.

The primary finding of this study was the inferior performance of both ChatGPT-4o and Microsoft Copilot in Arabic, with significantly lower ratings compared to the English responses. This aligns with recent research showing the challenges genAI models face in non-English languages, particularly in Arabic [30,32,33]. For example, Samaan *et al*., demonstrated that the accuracy of ChatGPT was lower in Arabic compared to English when addressing cirrhosis-related questions [51]. Recent studies also revealed the inferior performance of genAI in Arabic in queries related to infectious diseases [30], general health [32], and virology [33].

Beyond Arabic, this pattern extends to other non-English languages. For example, similar shortcomings have been reported in Chinese [31], Polish [52], and Spanish [53], further emphasizing the limitations of current genAI models in multilingual settings. These findings hint to the need for further development and fine-tuning of the currently available genAI models to improve performance in non-English languages, particularly Arabic. As genAI models are expected to play a larger role in healthcare, addressing these linguistic limitations is essential to ensure equitable access to accurate and reliable health information for

diverse global populations.

The language disparity observed in this study can be attributed to the fact that genAI models, including ChatGPT-4o and Microsoft Copilot, are primarily trained on English-based datasets. Despite efforts to incorporate more diverse languages, English continues to dominate both the quantity and quality of the digital content used for AI training [54]. As a result, even state-of-the-art gen-AI models struggle with languages that deviate significantly from English in their syntactic and grammatical structures [55]. Thus, it is conceivable that Arabic, with its complexity and varied dialects, remains underrepresented in AI training data, creating challenges for the equitable application of AI in healthcare [32]. This becomes particularly problematic when patients and providers require accurate, language-specific information to ensure effective treatment and self-management. The performance gap between English and other languages like Arabic has significant implications for patient self-help and health equity [56]. As AI-driven health information systems become more prevalent, especially for managing chronic conditions like diabetes and endocrine disorders, non-English-speaking populations may face difficulties in accessing high-quality medical information [57].

Our study identified the primary weaknesses in genAI models, regardless of language, which were the completeness and relevance of the content generated. To address these gaps, several key recommendations emerge. First, AI developers must prioritize integrating diverse, high-quality data for non-English languages, particularly those with complex structures like Arabic. Ongoing benchmarking across languages and medical fields is essential to close performance gaps. Second, healthcare providers should educate patients about the limitations of the currently available genAI tools, especially in non-English contexts, encouraging them to verify AI-generated information with trusted sources or consult professionals. Lastly, collaborating with multilingual medical experts in the development and validation of genAI models is crucial to ensure content is both clinically accurate and culturally appropriate. By addressing these issues, the accuracy, reliability, and equity of AI-driven health tools can be enhanced for the benefit of diverse populations.

This study has several limitations that future research should address as follows. First, the scope of the study was limited to 20 endocrine-related queries, restricting its generalizability to other medical fields. Second, the analysis focused only on English and Arabic, leaving the performance of genAI models in other languages unexplored. Third, real-world patient queries are more diverse and complex than the standardized questions used. Fourth, potential translation issues may have also affected the quality of Arabic responses, and expanding the study to include other medical specialties is necessary for broader applicability. Finally, future research should compare genAI-generated responses to those from human experts to establish clinical benchmarks.

CONCLUSIONS

To conclude, while ChatGPT-4o outperformed Microsoft Copilot across all CLEAR components, the observed language-based disparities highlight critical areas for improvement in the currently available genAI models. As these tools become more integral to healthcare delivery and patient education, it is vital to address these limitations to

prevent disadvantaging patients in non-English-speaking regions. Enhancing multilingual performance through improved training, fine-tuning, and validation will be essential to ensuring equitable healthcare access and supporting global patient self-help initiatives. This study highlighted the need for continued refinement of genAI models to ensure their effectiveness in diabetes care and other medical fields, regardless of language or region.

Author Contributions

Conceptualization, Hiba Abbasi and Muna Barakat; methodology, Hiba Abbasi, Marwa Al-Qudheeby, Zahra Ahmed Kheyami, Roaa Khalil, Nadia Khamees, Ola Hijjawi, Mohammed Sallam, Muna Barakat; validation, Hiba Abbasi and Muna Barakat; formal analysis, Hiba Abbasi, Roaa Khalil, Mohammed Sallam and Muna Barakat; investigation, Hiba Abbasi, Marwa Al-Qudheeby, Zahra Ahmed Kheyami, Roaa Khalil, Nadia Khamees, Ola Hijjawi, Mohammed Sallam, Muna Barakat; data curation, Hiba Abbasi, Marwa Al-Qudheeby, Zahra Ahmed Kheyami, Roaa Khalil, Nadia Khamees, Ola Hijjawi, Mohammed Sallam, Muna Barakat; writing—original draft preparation, Hiba Abbasi; writing—review and editing, Hiba Abbasi, Marwa Al-Qudheeby, Zahra Ahmed Kheyami, Roaa Khalil, Nadia Khamees, Ola Hijjawi, Mohammed Sallam, Muna Barakat; visualization, Hiba Abbasi and Muna Barakat; supervision, Hiba Abbasi and Muna Barakat; project administration, Hiba Abbasi. All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

The data presented in this study are available on request from the corresponding author (H.A.).

This manuscript is submitted in the special issue "**Evaluating Generative AI-Based Models in Healthcare**"

Acknowledgments: NA.

REFERENCES

1. Wu J, Lin X, Huang X, Shen Y, Shan PF. Global, regional and national burden of endocrine, metabolic, blood and immune disorders 1990-2019: a systematic analysis of the Global Burden of Disease study 2019. Front Endocrinol (Lausanne) 2023; 14: 1101627.
2. Hossain MJ, Al-Mamun M, Islam MR. Diabetes mellitus, the fastest growing global public health concern: Early detection should be focused. Health Sci Rep 2024; 7: e2004.
3. The L. Diabetes: a defining disease of the 21st century. Lancet 2023; 401: 2087.
4. Jia W. Diabetes: a challenge for China in the 21st century. Lancet Diabetes Endocrinol 2014; 2: e6-e7.
5. Kharroubi AT, Darwish HM. Diabetes mellitus: The epidemic of the century. World J Diabetes 2015; 6: 850-67.
6. Lin X, Xu Y, Pan X, et al. Global, regional, and national burden and trend of diabetes in 195 countries and territories: an analysis from 1990 to 2025. Sci Rep 2020; 10: 14790.
7. Ong KL, Stafford LK, McLaughlin SA, et al. Global, regional, and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: a systematic analysis for the Global Burden of Disease Study 2021. The Lancet 2023; 402: 203-34.
8. Chong B, Kong G, Shankar K, et al. The global syndemic of metabolic diseases in the young adult population: A consortium of trends and projections from the Global Burden of Disease 2000-2019. Metabolism 2023; 141: 155402.

9. El-Kebbi IM, Bidikian NH, Hneiny L, Nasrallah MP. Epidemiology of type 2 diabetes in the Middle East and North Africa: Challenges and call for action. World J Diabetes 2021; 12: 1401-25.
10. Clemente-Suárez VJ, Martín-Rodríguez A, Redondo-Flórez L, López-Mora C, Yáñez-Sepúlveda R, Tornero-Aguilera JF. New Insights and Potential Therapeutic Interventions in Metabolic Diseases. Int J Mol Sci 2023; 24: 10672.
11. Bungau AF, Tit DM, Bungau SG, et al. Exploring the Metabolic and Endocrine Preconditioning Associated with Thyroid Disorders: Risk Assessment and Association with Acne Severity. International Journal of Molecular Sciences 2024; 25: 721.
12. Jha BK, Sherpa ML, Imran M, et al. Progress in Understanding Metabolic Syndrome and Knowledge of Its Complex Pathophysiology. Diabetology 2023; 4: 134-59.
13. Chew NWS, Ng CH, Tan DJH, et al. The global burden of metabolic disease: Data from 2000 to 2019. Cell Metabolism 2023; 35: 414-28.e3.
14. Shriraam V, Mahadevan S, Anitharani M, Selvavinayagam, Sathiyasekaran B. National health programs in the field of endocrinology and metabolism - Miles to go. Indian J Endocrinol Metab 2014; 18: 7-12.
15. Karachaliou F, Simatos G, Simatou A. The Challenges in the Development of Diabetes Prevention and Care Models in Low-Income Settings. Front Endocrinol (Lausanne) 2020; 11: 518.
16. Eseadi C, Amedu AN, Ilechukwu LC, Ngwu MO, Ossai OV. Accessibility and utilization of healthcare services among diabetic patients: Is diabetes a poor man's ailment? World J Diabetes 2023; 14: 1493-501.
17. Sallam M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. Healthcare (Basel) 2023; 11: 887.
18. 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: 95-108.
19. Sallam M. Bibliometric top ten healthcare-related ChatGPT publications in the first ChatGPT anniversary. Narra J 2024; 4: e917.
20. 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 2023; 15: e35029.
21. Kanbach DK, Heiduk L, Blueher G, Schreiter M, Lahmann A. The GenAI is out of the bottle: generative artificial intelligence from a business model innovation perspective. Review of Managerial Science 2024; 18: 1189-220.
22. Shahsavar Y, Choudhury A. User Intentions to Use ChatGPT for Self-Diagnosis and Health-Related Purposes: Cross-sectional Survey Study. JMIR Hum Factors 2023; 10: e47564.
23. Schütz P, Lob S, Chahed H, et al. ChatGPT as an Information Source for Patients with Migraines: A Qualitative Case Study. Healthcare (Basel) 2024; 12: 1594.
24. Xu R, Wang Z. Generative artificial intelligence in healthcare from the perspective of digital media: Applications, opportunities and challenges. Heliyon 2024; 10: e32364.
25. Wang X, Sanders HM, Liu Y, et al. ChatGPT: promise and challenges for deployment in low- and middle-income countries. Lancet Reg Health West Pac 2023; 41: 100905.
26. Ernawati U, Wihastuti TA, Utami YW. Effectiveness of diabetes self-management education (DSME) in type 2 diabetes mellitus (T2DM) patients: Systematic literature review. J Public Health Res 2021; 10.
27. Bala I, Pindoo IA, Mijwil MM, 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 Medical Journal 2024; 58: 250-70.
28. Williamson SM, Prybutok V. The Era of Artificial Intelligence Deception: Unraveling the Complexities of False Realities and Emerging Threats of Misinformation. Information 2024; 15: 299.
29. 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 Pers Med 2024; 14: 568.
30. Sallam M, Al-Mahzoum K, Alshuaib O, et al. Language discrepancies in the performance of generative artificial intelligence models: an examination of infectious disease queries in English and Arabic. BMC Infectious Diseases 2024; 24: 799.
31. 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 2024; 26: e51926.
32. 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.
33. Sallam M, Al-Mahzoum K, Almutawaa RA, et al. The performance of OpenAI ChatGPT-4 and Google Gemini in virology multiple-choice questions: a comparative analysis of English and Arabic responses. BMC Research Notes 2024; 17: 247.
34. 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.
35. 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: e49373.
36. Sallam M, Al-Salahat K, Eid H, Egger J, Puladi B. Human versus Artificial Intelligence: ChatGPT-4 Outperforming Bing, Bard, ChatGPT-3.5, and Humans in Clinical Chemistry Multiple-Choice Questions. Adv Med Educ Pract 2024; 2024: 857-71.
37. Yilmaz Muluk S, Olcucu N. The Role of Artificial Intelligence in the Primary Prevention of Common Musculoskeletal Diseases. Cureus 2024; 16: e65372.
38. Sallam M, Al-Mahzoum K, Marzoaq O, et al. Evident gap between generative artificial intelligence as an academic editor compared to human editors in scientific publishing. Edelweiss Applied Science and Technology 2024; 8: 960-79.
39. Rosner BA. Fundamentals of biostatistics: Thomson-Brooks/Cole Belmont, CA, 2006.
40. Mohd Arof K, Ismail S, Saleh AL. Contractor's Performance Appraisal System in the Malaysian Construction Industry: Current Practice, Perception and Understanding. International Journal of Engineering & Technology 2018; 7: 46.
41. Maina PM, Pienaar M, Reid M. Self-management practices for preventing complications of type II diabetes mellitus in low and middle-income countries: A scoping review. Int J Nurs Stud Adv 2023; 5: 100136.
42. 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; 3: 100105.
43. Strzelecki A. Is ChatGPT-like technology going to replace commercial search engines? Library Hi Tech News 2024; 41: 18-21.
44. Alanezi F. Factors influencing patients' engagement with ChatGPT for accessing health-related information. Critical Public Health 2024; 34: 1-20.
45. 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; 16: e63580.

46. Podder I, Pipil N, Dhabal A, Mondal S, Pienyii V, Mondal H. Evaluation of Artificial Intelligence-Based Chatbot Responses to Common Dermatological Queries. Jordan Medical Journal 2024; 58: 271–8.
47. Kaftan AN, Hussain MK, Naser FH. Response accuracy of ChatGPT 3.5 Copilot and Gemini in interpreting biochemical laboratory data a pilot study. Scientific Reports 2024; 14: 8233.
48. Mayo-Yáñez M, Lechien JR, Maria-Saibene A, Vaira LA, Maniaci A, Chiesa-Estomba CM. Examining the Performance of ChatGPT 3.5 and Microsoft Copilot in Otolaryngology: A Comparative Study with Otolaryngologists' Evaluation. Indian Journal of Otolaryngology and Head & Neck Surgery 2024; 76: 3465-9.
49. Sallam M, Al-Salahat K, Al-Ajlouni E. ChatGPT Performance in Diagnostic Clinical Microbiology Laboratory-Oriented Case Scenarios. Cureus 2023; 15: e50629.
50. Sallam M, Khalil R, Sallam M. Benchmarking Generative AI: A Call for Establishing a Comprehensive Framework and a Generative AIQ Test. Mesopotamian Journal of Artificial Intelligence in Healthcare 2024; 2024: 69-75.
51. Samaan JS, Yeo YH, Ng WH, et al. ChatGPT's ability to comprehend and answer cirrhosis related questions in Arabic. Arab J Gastroenterol 2023; 24: 145-8.
52. 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. Sci Rep 2023; 13: 20512.
53. Guillen-Grima F, Guillen-Aguinaga S, Guillen-Aguinaga L, 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: 1460-87.
54. Tang H-WV, Yin M-S, Sheu R-S. The relationship between English language adoption and global digital inequality: A cross-country analysis of ICT readiness and use. In: System and technology Advancements in Distance Learning: IGI Global, 2013; 124-33.
55. Nicholas G, Bhatia A. Lost in translation: large language models in non-English content analysis. arXiv preprint arXiv:2306.07377 2023.
56. Al Shamsi H, Almutairi AG, Al Mashrafi S, Al Kalbani T. Implications of Language Barriers for Healthcare: A Systematic Review. Oman Med J 2020; 35: e122.
57. Pelicioni PHS, Michell A, Santos P, Schulz JS. Facilitating Access to Current, Evidence-Based Health Information for Non-English Speakers. Healthcare (Basel) 2023; 11: 1932.

تقييم عابر للغات لنماذج الذكاء الاصطناعي التوليدي لتساؤلات متعلقة بأمراض السكري والغدد الصماء

هبة العباسي[1،2]، مروة القضيبي[3]، زهرة أحمد خيامي[4]، رؤى خليل[5]، ناديا خميس[1،2]،

علا الحجاوي[1،2]، محمد سلّام[6]، منى بركات[7]

[1] قسم الأمراض الباطنية، كلية الطب، الجامعة الأردنية، عمان، الأردن

[2] دائرة الأمراض الباطنية، مستشفى الجامعة الأردنية، عمان، الأردن

[3] وزارة الصحة الكويتية، مستشفى مبارك الكبير، الجابرية، الكويت.

[4] وزارة الصحة الكويتية ، الكويت

[5] قسم علم الأمراض والأحياء الدقيقة والطب الشرعي، كلية الطب، الجامعة الأردنية، عمان، الأردن

[6] قسم الصيدلة، مستشفى ميديكلينيك بارك فيو، ميديكلينيك الشرق الأوسط، دبي، الإمارات العربية المتحدة

[7] قسم الصيدلة السريرية والعلاجية، كلية الصيدلة، جامعة العلوم التطبيقية الخاصة، عمان، الأردن

Received: September 18, 2024
Accepted: October 22, 2024

DOI:
https://doi.org/10.35516/jmj.v58i4.3369

الملخص

الخلفية والأهداف: تشكل اضطرابات الغدد الصماء والأيض، بما في ذلك مرض السكري، تحديات صحية عالمية كبرى. تُستخدم نماذج الذكاء الاصطناعي التوليدي بشكل متزايد للمساعدة الذاتية للمرضى. تهدف هذه الدراسة إلى تقييم أداء نموذجي الذكاء الاصطناعي التوليدي، ChatGPT وMicrosoft Copilot، في الاستفسارات المتعلقة بالغدد الصماء باللغتين الإنجليزية والعربية.

المواد والطرق: التزمت هذه الدراسة الوصفية بقائمة METRICS لدراسات الذكاء الصناعي التوليدي في مجال الرعاية الصحية، حيث تمت مقارنة الاستجابات من ChatGPT-4o وMicrosoft Copilot باستخدام 20 استعلامًا متعلقًا بالغدد الصماء باللغتين الإنجليزية والعربية. تم تقييم الاستجابات باستخدام أداة CLEAR، التي قامت بتقييم الاكتمال والدقة والملاءمة. قام ثلاثة خبراء في الغدد الصماء بتقييم مخرجات الذكاء الصناعي التوليدي بشكل مستقل.

النتائج: تم تقييم إجمالي 80 استجابة. كانت موثوقية التقييم بين المقيمين عالية (Cronbach α = 0.832، معامل الارتباط داخل الفئة = 0.832). تفوق ChatGPT-4o باستمرار على Microsoft Copilot، وحصل على تصنيفات "ممتاز" باللغة الإنجليزية و"جيد جدًا" باللغة العربية، بينما حقق Microsoft Copilot تصنيف "جيد جدًا" باللغة الإنجليزية و"جيد" إلى "جيد جدًا" باللغة العربية. تفوق ChatGPT-4o على Microsoft Copilot في الاكتمال (4.38 مقابل 3.36، p <.001)، والدقة (4.18 مقابل 3.83، p = .014)، والملاءمة (4.44 مقابل 3.82، p <.001). وقد تباين الأداء بشكل كبير بين الإجابات الإنجليزية والعربية، حيث بلغت قيمة p<.001 للاكتمال، وقيمة p=.001 للدقة، وقيمة p=.012 للأهمية، وقيمة p<.001 للنتيجة الإجمالية لـ CLEAR. ولم يتم العثور على فروق ذات دلالة إحصائية بناءً على موضوع الاستعلام.

الاستنتاجات: لقد تفوق برنامج ChatGPT-4o على برنامج Microsoft Copilot في جميع مكونات CLEAR، ولكن كانت هناك فجوات واضحة فيما يتعلق باللغة. إن معالجة هذه القيود أمر بالغ الأهمية لضمان الوصول العادل للمرضى غير الناطقين باللغة الإنجليزية.

الكلمات الدالة: الذكاء الاصطناعي؛ المحولات المدربة مسبقًا؛ معالجة اللغة الطبيعية؛ ممارسة الرعاية الصحية؛ مرض السكري.