Medical Journal of Bakırköy

Med J Bakirkoy 2025;21(3):310-315
DOI: 10.4274/BMJ.galenos.2025.2025.5-9

Research

Evaluation of ChatGPT-4o's Responses to Questions about Myasthenia Gravis in English and Turkish

ChatGPT-4o'nun Miyastenia Gravis Hakkındaki İngilizce ve Türkçe Sorulara Verdiği Yanıtların Değerlendirilmesi

Berin İnan, Ömer Karadaş, Zeki Odabaşı

University of Health Sciences Türkiye, Gülhane Faculty of Medicine, Department of Neurology, Ankara, Türkiye

ABSTRACT

Objective: Large language models, such as Chat Generative Pre-Trained Transformer 4o (ChatGPT-4o), are increasingly used by both patients and medical professionals to access health-related information. Myasthenia gravis (MG) is a chronic autoimmune neuromuscular disorder requiring long-term treatment. Therefore, timely access to accurate medical information about MG is important. This study aimed to evaluate the accuracy, completeness, clarity, appropriateness for the target audience, risk of misinformation or harm, and readability of ChatGPT-4o-generated responses to queries about MG from patients and neurology residents, in both English and Turkish.

Methods: We developed four sets of 20 questions, frequently asked by patients and neurology residents about MG in both English and Turkish, covering pathophysiology and symptoms, diagnosis, treatment, prognosis, and daily management. ChatGPT-4o responses were generated in separate sessions on March 29, 2025. Two neurologists independently evaluated the responses using a 5-point Likert scale across five domains. Readability was assessed using the Flesch Reading Ease score, Flesch-Kincaid grade level, and Gunning-Fog index for English, and the Ateşman readability index for Turkish.

Results: Scores for accuracy, clarity, appropriateness, and risk of misinformation or harm were consistently above 4 in both languages, with clarity rated as 5 in all responses. Completeness received the lowest scores (3.5-5.0), particularly in Turkish responses to resident-directed questions. Readability was higher in Turkish. English responses to resident queries were extremely difficult to read, while patient-directed ones remained in the "difficult" to "very difficult" range. Several discrepancies were observed in specific contents between English and Turkish outputs, such as differences in differential diagnosis lists, treatment options, contraindicated medications, and thymectomy indications.

Conclusion: ChatGPT-4o produced high-quality responses overall to MG-related queries in both languages. However, language-specific inconsistencies and content omissions highlight the need for further model refinement, particularly in multilingual and professional-use contexts.

Keywords: Artificial intelligence, ChatGPT-4o, large language models, myasthenia gravis, neurology

ÖZ

Amaç: Chat Generative Pre-Trained Transformer 4o (ChatGPT-4o) gibi büyük dil modelleri, hem hastalar hem de sağlık profesyonelleri tarafından sağlıkla ilgili bilgilere erişmek amacıyla giderek daha fazla kullanılmaktadır. Miyastenia gravis (MG), uzun süreli tedavi gerektiren kronik bir otoimmün nöromüsküler hastalıktır. Bu nedenle, MG hakkında doğru tıbbi bilgilere zamanında erişim önemlidir. Bu çalışma, ChatGPT-4o tarafından MG ile ilgili olarak hastalar ve nöroloji asistanları tarafından yöneltilen İngilizce ve Türkçe sorulara verilen yanıtların doğruluk, bütünlük, açıklık, hedef kitleye uygunluk, yanlış bilgilendirme veya zarar riski ile okunabilirlik açısından değerlendirilmesini amaçlamıştır.

Gereç ve Yöntem: İngilizce ve Türkçe olarak, hastalar ve nöroloji asistanları tarafından MG hakkında sıkça sorulan sorulardan oluşan, patofizyoloji ve semptomlar, tanı, tedavi, prognoz ve günlük yaşam yönetimini kapsayan 20 soruluk dört soru seti oluşturuldu. Her bir set için ChatGPT-4o yanıtları, 29 Mart 2025 tarihinde ayrı oturumlarda üretildi. İki nörolog, yanıtları beş farklı alanda 5 puanlık Likert ölçeği kullanarak birbirinden bağımsız olarak değerlendirdi. Okunabilirlik; İngilizce için Flesch okuma kolaylığı skoru, Flesch-Kincaid sınıf düzeyi ve Gunning-Fog indeksi ile, Türkçe için ise Ateşman okunabilirlik indeksi ile değerlendirildi.

Bulgular: Doğruluk, açıklık, uygunluk ve yanlış bilgilendirme ya da zarar verme riski açısından puanlar her iki dilde de tutarlı şekilde 4'ün üzerinde olup, açıklık tüm yanıtlarda 5 olarak değerlendirilmiştir. Bütünlük ise, özellikle asistanlara yönelik Türkçe yanıtlarda en düşük puanları (3,5-5,0) almıştır. Okunabilirlik Türkçe'de daha yüksek bulunmuştur. Asistanlara yönelik İngilizce yanıtlar son derece zor okunabilirken, hastalara yönelik

Address for Correspondence: Berin İnan, MD, University of Health Sciences Türkiye, Gülhane Faculty of Medicine, Department of Neurology, Ankara, Türkiye
E-mail: berin.inan@yahoo.com **ORCID ID:** orcid.org/0000-0002-2758-0207

Cite as: İnan B, Karadaş Ö, Odabaşı Z. Evaluation of ChatGPT-4o's responses to questions about myasthenia gravis in English and Turkish. Med J Bakirkoy. 2025;21(3):310-315

Received: 22.05.2025
Accepted: 08.07.2025
Publication Date: 03.09.2025

olanlar "zor" ile "çok zor" arasında değişmiştir. İngilizce ve Türkçe yanıtlar arasında; ayırıcı tanı listesi, tedavi seçenekleri, kontrendike ilaçlar ve timektomi endikasyonları gibi belirli içeriklerde çeşitli tutarsızlıklar gözlemlenmiştir.

Sonuç: ChatGPT-4o, MG ile ilgili sorulara her iki dilde de genel olarak yüksek kaliteli yanıtlar üretmiştir. Ancak dile özgü tutarsızlıklar ve içerik eksiklikleri, özellikle çok dilli ve profesyonel kullanım bağlamlarında modelin daha da geliştirilmesi gerektiğini ortaya koymaktadır.

Anahtar Kelimeler: Yapay zeka, ChatGPT-4o, büyük dil modelleri, miyastenia gravis, nöroloji

INTRODUCTION

In recent years, artificial intelligence (AI) has been increasingly used in the medical field by both patients and healthcare professionals. In particular, large language models (LLMs), a type of machine learning model designed to understand, analyze, generate, and manipulate human language, have gained popularity as fast, easy, and accessible tools for addressing a broad range of inquiries, from everyday concerns to complex academic questions (1,2).

Myasthenia gravis (MG) is the most common disorder of the neuromuscular junction and is characterized by fatigable skeletal muscle weakness (3,4). As a chronic autoimmune condition that necessitates long-term treatment, MG significantly impacts patients' quality of life and requires effective patient education and engagement (5). Patients frequently use the internet to search for information about the signs and symptoms of the disease, available treatment options, and strategies for the daily management of myasthenic symptoms (6,7). Similarly, medical professionals increasingly rely on LLMs for rapid access to information about a variety of medical conditions, including MG (6,7). However, the accuracy and reliability of the AI-generated content vary, requiring careful evaluation, as LLMs may have a tendency to hallucinate, resulting in misinformation (8-11).

Chat Generative Pre-Trained Transformer 4 (ChatGPT-4), one of the widely used LLMs, was launched by OpenAI in March 2023 (1). ChatGPT-4 is capable of processing both text and image inputs and performing complex tasks (1). More recently, an advanced version, ChatGPT-4o, was released by OpenAI in May 2024 (12). ChatGPT-4o is superior to ChatGPT-4 in terms of speed, cost-efficiency, multimodal functionality, and multilingual performance (2,12,13). While ChatGPT-4o is available for free with usage limitations and also as a paid version with extended features, ChatGPT-4 is not freely accessible (13).

In this study, we aimed to evaluate and compare the accuracy, completeness, clarity, appropriateness for the target audience, risk of misinformation or harm, and readability of ChatGPT-4o-generated responses to queries about MG from patients and neurology residents, in both English and Turkish.

METHODS

Study Design and Analysis of Responses

In this cross-sectional study, we developed four sets of queries consisting of frequently asked questions about MG from patients and neurology residents, in both English and Turkish. The questions were identical across the English and Turkish versions, and the content of the questions was similar in the patient and neurology resident groups. However, while the questions were phrased in a scientific tone in the neurology resident group, plain language was preferred for the patient group. Each set of queries consisted of 20 questions, including five on pathophysiology and symptoms, three on diagnosis, four on treatment modalities, four on prognosis, and four on the daily management of MG.

Each set of queries was submitted to ChatGPT-4o separately in a new chat window on 29 March 2025. The responses generated by ChatGPT-4o were independently evaluated by two neurologists specialized in neuromuscular diseases, based on the current literature about MG (4). Evaluations were conducted across five domains: accuracy, completeness, clarity, appropriateness for the target audience, and risk of misinformation or harm. Each domain was assessed using a 5-point Likert scale, ranging from 1 (poor) to 5 (excellent). The mean of the scores given by the two experts for each domain was calculated and used for statistical analysis.

We analyzed the readability of the responses in English using Readable software (Readable.com, Horsham, United Kingdom) (14), applying the Flesch Reading Ease score (FRES), Flesch-Kincaid grade level, and Gunning-Fog index. The readability of the responses in Turkish was evaluated using the Ateşman readability index (15,16).

Statistical Analysis

The normality of the data was assessed using the Shapiro-Wilk test. Descriptive statistics were given as mean±standard deviation or median (minimum-maximum) for continuous variables, and as frequency (percentage) for categorical variables. Group comparisons were performed using the Mann-Whitney U test or the independent samples t-test. All statistical analyses were conducted using SPSS for Windows, version 23.0 (IBM Corp., Armonk, NY, USA). Statistical significance was set as a p-value <0.05.

Ethical Approval

This cross-sectional study did not involve human participants, human tissue, or individually identifiable data. Therefore, informed consent and ethical approval were waived in accordance with institutional and national guidelines.

RESULTS

The accuracy, clarity, appropriateness for the target audience, and low risk of misinformation or harm scores for all ChatGPT-4o-generated responses were above 4 (very good) in both Turkish and English. All responses received a score of 5 points (excellent) for clarity from both experts. Among all evaluated domains, completeness received the lowest scores, ranging from 3.5 to 5 points. Although statistical analysis was not performed for the subgroups (pathophysiology and symptoms, diagnosis, treatment modalities, prognosis, and daily life management) due to the small sample size, we observed that completeness scores were higher for responses related to prognosis and daily life management than for those addressing pathophysiology and symptoms, diagnosis, and treatment modalities.

The scores for accuracy, completeness, clarity, appropriateness for the target audience, and risk of misinformation or harm were comparable between the patient and resident groups in the English responses. However, in the Turkish responses, the patient group demonstrated higher completeness scores than the resident group (Tables 1 and 2).

In the assessment of readability in English, responses to the patient group were found to be easier to read (Table 1). Readability scores were similar between the patient and neurology resident groups in Turkish responses (Table 2).

When comparing English and Turkish responses, the scores for accuracy, completeness, clarity, appropriateness, and risk of misinformation or harm did not differ in responses to patient queries. However, the completeness scores for the resident group were significantly lower in Turkish than in English (Table 3).

Table 1. Analysis of ChatGPT-4o's responses in English

Parameters	For patients	For residents	p-value
Accuracy, median (min-max)	5.0 (4.0-5.0)	5.0 (4.0-5.0)	0.799
Completeness, median (min-max)	5.0 (3.5-5.0)	5.0 (4.0-5.0)	0.602
Clarity, median (min-max)	5.0 (5.0-5.0)	5.0 (5.0-5.0)	1.000
Appropriateness for audience, median (min-max)	5.0 (5.0-5.0)	5.0 (4.0-5.0)	0.799
Risk of misinformation or harm, median (min-max)	5.0 (4.5-5.0)	5.0 (5.0-5.0)	0.799
Readability scores			
FRES, mean±SD	28.1±22.6	1.0±26.4	**0.001**
FKGL, mean±SD	11.5±3.5	15.2±3.9	**0.003**
GFI, mean±SD	12.0±3.4	14.9±4.3	**0.024**

FKGL: Flesch-Kincaid grade level, FRES: Flesch Reading Ease score, GFI: Gunning fog index, SD: Standard deviation

Table 2. Analysis of ChatGPT-4o's responses in Turkish

Parameters	For patients	For residents	p-value
Accuracy, median (min-max)	5.0 (4.5-5.0)	5.0 (4.0-5.0)	0.799
Completeness, median (min-max)	5.0 (3.5-5.0)	5.0 (4.0-5.0)	0.602
Clarity, median (min-max)	5.0 (5.0-5.0)	5.0 (5.0-5.0)	1.000
Appropriateness for audience, median (min-max)	5.0 (4.5-5.0)	5.0 (4.0-5.0)	0.799
Risk of misinformation or harm, median (min-max)	5.0 (5.0-5.0)	5.0 (5.0-5.0)	0.799
Ateşman readability index, median (min-max)	69.9 (40.2-91.2)	78.2 (25.3-95.7)	0.068

Table 3. Comparison of ChatGPT-4o's responses in English and Turkish

	English	Turkish	p-value
For patients			
Accuracy, median (min-max)	5.0 (4.0-5.0)	5.0 (4.5-5.0)	0.968
Completeness, median (min-max)	5.0 (3.5-5.0)	5.0 (3.5-5.0)	0.620
Clarity, median (min-max)	5.0 (5.0-5.0)	5.0 (5.0-5.0)	1.000
Appropriateness for audience, median (min-max)	5.0 (5.0-5.0)	5.0 (4.5-5.0)	0.799
Risk of misinformation or harm, median (min-max)	5.0 (4.5-5.0)	5.0 (5.0-5.0)	0.799
For residents			
Accuracy, median (min-max)	5.0 (4.0-5.0)	5.0 (4.5-5.0)	0.904
Completeness, median (min-max)	5.0 (4.0-5.0)	4.5 (3.5-5.0)	**0.035**
Clarity, median (min-max)	5.0 (5.0-5.0)	5.0 (5.0-5.0)	1.000
Appropriateness for audience, median (min-max)	5.0 (4.0-5.0)	5.0 (4.0-5.0)	0.192
Risk of misinformation or harm, median (min-max)	5.0 (5.0-5.0)	5.0 (5.0-5.0)	1.000

DISCUSSION

In this study, we examined the performance of ChatGPT-4o in responding to frequently asked queries about MG, both in English and Turkish, by patients, and neurology residents. Rather than focusing solely on quantitative scores, we also analyzed qualitative aspects such as clarity, completeness of the content, contextual appropriateness for the target audiences, and linguistic differences.

ChatGPT-4o exhibited high overall response quality and reliability in our study, as reflected by high accuracy, clarity, appropriateness scores, and low risk of misinformation. However, the performance of ChatGPT models has shown variability depending on the complexity of content and disease subtypes (17-20). ChatGPT-3.5 achieved stronger results in peripheral nerve and cerebrovascular diseases, while its performance was weaker in neuromuscular junction disorders and multiple sclerosis (17). Although ChatGPT-4's performance matched or even surpassed physicians in multiple-choice and board-style exams (18,19), its capacity for clinical reasoning and higher-order decision making remained limited (18). Furthermore, concerns regarding overconfidence and factual inconsistency have been raised, particularly in open-ended questions or complex medical scenarios (18-20).

The completeness scores were the lowest among all domains, which exhibited a language-dependent variation, prompting further evaluation. Completeness scores were notably lower in Turkish responses to resident queries compared to their English counterparts. This finding is consistent with prior studies suggesting that ChatGPT-4 exhibits higher performance in English, likely due to the predominance of English-language training data in the model's training corpus (21-24). Studies evaluating the performance of ChatGPT-4 in bilingual examinations showed that the model achieved significantly higher accuracy in English than in Chinese (21), Arabic (23), and Korean (24). Language-related discrepancies in ChatGPT-4 performance have also been reported in clinical and public health domains. In a study evaluating ChatGPT's multilingual performance in clinical nutrition advice, ChatGPT-4 produced significantly lower quality outputs in Kazakh compared to English and Russian (22). These findings collectively suggest that language-based performance biases remain a challenge and underscore the need for multilingual fine-tuning and dataset diversification to address such disparities. Although ChatGPT-4o has been promoted as a multilingually improved version of its predecessors (2,12), to the best of our knowledge, no study has yet evaluated its multilingual performance in any medical context, including MG.

In addition to the differences in completeness scores between English and Turkish responses, our study also identified linguistic discrepancies in specific content elements. For instance, in response to the patient group's question, "Can MG be mistaken for other conditions?", the differential diagnoses listed differed between the English and Turkish versions. Similarly, the response to the question on treatment options included eculizumab in English but omitted it in Turkish. The list of contraindicated medications in MG also varied between the two languages. Similar discrepancies were observed in resident-directed queries. The differential diagnoses of MG were not consistent across languages, and the indications for thymectomy included different age thresholds in the English and Turkish responses. Moreover, the predictors of spontaneous remission varied

between languages. These findings suggest potential inconsistencies in how ChatGPT-4o retrieves and generates language-specific medical content.

Notably, in our study, ChatGPT-4o achieved its highest completeness scores when responding to questions related to prognosis and the daily management of MG, suggesting the model's strength in patient-centered communication. This finding is consistent with previous studies across various medical conditions (9,11,25). ChatGPT-4o has been shown to provide accurate and reliable responses in contexts such as keratoconus (25), prostate cancer (9), and postmenopausal osteoporosis (11), particularly when addressing follow-up care, treatment adherence, and lifestyle recommendations.

Readability remains a significant barrier to patient accessibility in AI-generated medical content. In our study, responses to both patient and resident-directed questions in Turkish were fairly easy to read. However, in their English counterparts, responses to resident-directed questions were extremely difficult to read based on the FRES, whereas responses to patient-directed questions showed higher scores, indicating relatively better accessibility. Nevertheless, the overall readability for responses to patient-directed questions remained within the "difficult" to "very difficult" range, aligning with previous studies that have highlighted the limited readability of ChatGPT-4o's outputs in various clinical contexts (8-11,25). Encouragingly, several studies have shown that prompting ChatGPT-4o to simplify its language can significantly improve readability without compromising accuracy (9,10). Moreover, patient perceptions of understandability may not always correspond to objective readability indices, suggesting that future models should integrate real-time feedback and personalization to improve communication effectiveness (9).

Study Limitations

This study has several limitations. First, the number of questions was limited, restricting the application of statistical analyses in certain comparisons. Second, the evaluation was based on predefined questions, which may not fully reflect the variability of real-world questions posed by patients or residents. Additionally, only two languages were assessed, limiting the generalizability of the findings to other languages. Finally, readability was evaluated using standard indices that may not accurately represent actual patient comprehension.

CONCLUSION

ChatGPT-4o demonstrated high overall performance in responding to medical queries about MG, providing accurate, clear, and contextually appropriate answers in both English and Turkish. Although minor language-related differences were observed, particularly in the completeness of complex responses and certain factual discrepancies, ChatGPT-4o shows strong potential as a supportive tool for both patient education and professional reference.

ETHICS

Ethics Committee Approval: This cross-sectional study did not involve human participants, human tissue, or individually identifiable data. Therefore, ethical approval were waived in accordance with institutional and national guidelines.

Informed Consent: Informed consent was not obtained.

FOOTNOTES

Authorship Contributions

Concept: B.İ., Z.O., Design: B.İ., Z.O., Data Collection or Processing: B.İ., Ö.K., Analysis or Interpretation: B.İ., Ö.K., Z.O., Literature Search: B.İ., Ö.K., Writing: B.İ.

Conflict of Interest: No conflict of interest was declared by the authors.

Financial Disclosure: The authors declare that this study received no financial support.

REFERENCES

1. OpenAI. GPT-4 technical report. arXiv Preprint. 2023; arXiv:2303.08774. Available from: https://arxiv.org/pdf/2303.08774
2. Shahriar S, Lund BD, Mannuru NR, Arshad MA, Hayawi K, Bevara RVK, et al. Putting GPT-4o to the sword: a comprehensive evaluation of language, vision, speech, and multimodal proficiency. Appl Sci. 2024;14:7782.
3. Gilhus NE, Tzartos S, Evoli A, Palace J, Burns TM, Verschuuren JJGM. Myasthenia gravis. Nat Rev Dis Primers. 2019;5:30.
4. Bird SJ. Myasthenia gravis [Internet]. UpToDate; 2025 [cited 2025 Mar 29]. Available from: https://www.uptodate.com/contents/overview-of-the-treatment-of-myasthenia-gravis?search=Overview%20of%20the%20treatment%20of%20myasthenia%20gravis&source=search_result&selectedTitle=1~150&usage_type=default&display_rank=1
5. Dewilde S, Philips G, Paci S, Beauchamp J, Chiroli S, Quinn C, et al. Patient-reported burden of myasthenia gravis: baseline results of the international prospective, observational, longitudinal real-world digital study MyRealWorld-MG. BMJ Open. 2023;13:e066445.
6. Jin JQ, Dobry AS. ChatGPT for healthcare providers and patients: practical implications within dermatology. J Am Acad Dermatol. 2023;89:870-1.

7. Tan S, Xin X, Wu D. ChatGPT in medicine: prospects and challenges: a review article. Int J Surg. 2024;110:3701-6.
8. Wang S, Wang Y, Jiang L, Chang Y, Zhang S, Zhao K, et al. Assessing the clinical support capabilities of ChatGPT 4o and ChatGPT 4o mini in managing lumbar disc herniation. Eur J Med Res. 2025;30:45.
9. Trapp C, Schmidt-Hegemann N, Keilholz M, Brose SF, Marschner SN, Schonecker S, et al. Patient- and clinician-based evaluation of large language models for patient education in prostate cancer radiotherapy. Strahlenther Onkol. 2025;201:333-42.
10. Li J, Chang C, Li Y, Cui S, Yuan F, Li Z, et al. Large language models' responses to spinal cord injury: a comparative study of performance. J Med Syst. 2025;49:39.
11. Liu R, Liu J, Yang J, Sun Z, Yan H. Comparative analysis of ChatGPT-4o mini, ChatGPT-4o and Gemini advanced in the treatment of postmenopausal osteoporosis. BMC Musculoskelet Disord. 2025;26:369.
12. OpenAI. GPT-4o System Card [Internet]. 2025 [cited 2025 Aug 8]. Available from: https://arxiv.org/pdf/2410.21276
13. Craig L. GPT-4o vs. GPT-4: How do they compare? [Internet]. 2025 [cited 2025]. Available from: https://www.techtarget.com/searchenterpriseai/feature/GPT-4o-vs-GPT-4-How-do-they-compare#:~:text=OpenAI's%20testing%20indicates%20that%20GPT,idioms%2C%20metaphors%20and%20cultural%20references.
14. Readable.com. Readability score [Internet]. [cited 2025]. Available from: https://app.readable.com/text/?demo&_ga=2.14124151.156675892.1746127328-404057203.1743896947
15. Ateşman E. Measurement of readability in Turkish. Language J [Internet]. 1997;58:71-4.
16. Türkçe okunabilirlik indeksi; 2025. Erişim adresi: http://okunabilirlikindeksi.com/
17. Altunisik E, Firat YE, Cengiz EK, Comruk GB. Artificial intelligence performance in clinical neurology queries: the ChatGPT model. Neurol Res. 2024;46:437-43.
18. Chen TC, Multala E, Kearns P, Delashaw J, Dumont A, Maraganore D, et al. Assessment of ChatGPT's performance on neurology written board examination questions. BMJ Neurol Open. 2023;5:e000530.
19. Ros-Arlanzón P, Perez-Sempere A. Evaluating AI competence in specialized medicine: comparative analysis of ChatGPT and neurologists in a neurology specialist examination in Spain. JMIR Med Educ. 2024;10:e56762.
20. García-Rudolph A, Sanchez-Pinsach D, Opisso E, Soler MD. Exploring new educational approaches in neuropathic pain: assessing accuracy and consistency of artificial intelligence responses from GPT-3.5 and GPT-4. Pain Med. 2025;26:48-50.
21. Wu Z, Gan W, Xue Z, Ni Z, Zheng X, Zhang Y. Performance of ChatGPT on nursing licensure examinations in the United States and China: cross-sectional study. JMIR Med Educ. 2024;10:e52746.
22. Adilmetova G, Nassyrov R, Meyerbekova A, Karabay A, Varol HA, Chan MY. Evaluating ChatGPT's multilingual performance in clinical nutrition advice using synthetic medical text: insights from central Asia. J Nutr. 2025;155:729-35.
23. Sallam M, Al-Mahzoum K, Alshuaib O, Alhajri H, Alotaibi F, Alkhurainej D, et al. Language discrepancies in the performance of generative artificial intelligence models: an examination of infectious disease queries in English and Arabic. BMC Infect Dis. 2024;24:799.
24. Song ES, Lee SP. Comparative analysis of the response accuracies of large language models in the Korean national dental hygienist examination across Korean and English questions. Int J Dent Hyg. 2025;23:267-76.
25. Balci AS, Cakmak S. Evaluating the accuracy and readability of ChatGPT-4o's responses to patient-based questions about keratoconus. Ophthalmic Epidemiol. 2025:1-6.