ARAŞTIRMA MAKALESİ / *RESEARCH ARTICLE*

Evaluating the Effectiveness of Artificial Intelligence Language Models in Hearing Loss Assessment: Comparative Study of ChatGPT, Gemini, and Perplexity

İşitme Kaybı Değerlendirmesinde Yapay Zeka Dil Modellerinin Etkinliğinin Değerlendirilmesi: ChatGPT, Gemini ve Perplexity'nin Karşılaştırmalı Bir Çalışması

Mehmet Zeki Erdem[1], Abdulaziz Yalinkilic[1], Yaser Said Cetin[1], Nizamettin Erdem[2]

[1]Department of Otorhinolaryngology; [2]Department of Family Medicine, Van Yuzuncu Yil University, Faculty of Medicine, Van, Türkiye

ABSTRACT

Aim: *Hearing loss constitutes a considerable health concern that adversely impacts individuals' quality of life. This study aims to compare the efficacy of artificial intelligence (AI) based methods in evaluating hearing loss.*

Material and Methods: *The study assessed the performance of the ChatGPT, Gemini, and Perplexity programs in terms of quality (GQS), accuracy (Likert), and readability (GFI). The data were analyzed utilizing nonparametric tests, and the intergroup differences were assessed with the Kruskal-Wallis H test. Groups exhibiting substantial differences were analyzed using the Bonferroni-adjusted Post-Hoc test.*

Results: *The investigation indicates that ChatGPT outperforms other tools in quality (p=0.018). No substantial differences were observed between the groups in accuracy (p=0.072) or readability (p>0.05). The GQS scores for ChatGPT, Gemini, and Perplexity were 4.93, 4.71, and 4.43, respectively.*

Conclusions: *ChatGPT has demonstrated improved quality in assessing hearing loss. Nonetheless, comparable outcomes regarding accuracy and readability indicate that alternative approaches may also prove effective in some applications. These findings endorse the efficacy of AI-based technologies for specific health concerns, such as evaluating hearing loss. In the future, it is advisable to evaluate these technologies using larger samples and across a range of health conditions.*

Key words: *hearing loss; artificial intelligence; chatGPT; gemini; perplexity*

ÖZET

Amaç: *İşitme kaybı, bireylerin yaşam kalitesini olumsuz etkileyen önemli bir sağlık sorunudur. Bu çalışmanın amacı, işitme kaybı değerlendirmesinde kullanılan yapay zekâ tabanlı araçların etkinliğini karşılaştırmaktır.*

Materyal ve Metod: *Çalışmada, ChatGPT, Gemini ve Perplexity programlarının kalite (GKS), doğruluk (Likert) ve okunabilirlik (GFI) açısından performansları değerlendirilmiştir. Veriler, nonparametrik testler kullanılarak analiz edilmiş ve gruplar arasındaki farklar Kruskal-Wallis H testi ile incelenmiştir. Anlamlı fark bulunan gruplar, Bonferroni düzeltmeli post hoc testi ile karşılaştırılmıştır.*

Bulgular: *Analiz sonuçları, ChatGPT'nin kalite açısından diğer araçlara göre üstün olduğunu göstermiştir (p=0,018). Ancak, doğruluk (p=0,072) ve okunabilirlik (p>0,05) açısından gruplar arasında anlamlı bir fark bulunmamıştır. ChatGPT'nin GKS puanı 4,93, Gemini'nin 4,71 ve Perplexity'nin 4,43 olarak hesaplanmıştır.*

Sonuç: *ChatGPT işitme kaybı değerlendirmesinde kalite açısından üstün performans sergilemiştir. Ancak, doğruluk ve okunabilirlik açısından benzer sonuçlar, diğer araçların da belirli kullanım alanlarında etkili olabileceğini göstermektedir. Bu bulgular, yapay zekâ tabanlı araçların işitme kaybı değerlendirmesi gibi spesifik sağlık sorunlarında kullanılabilirliğini desteklemektedir. Gelecekte, bu araçların daha geniş örneklemlerle ve farklı sağlık sorunlarında test edilmesi önerilmektedir.*

Anahtar kelimeler: *işitme kaybı; yapay zekâ; chatGPT; gemini; perplexity*

İletişim/Contact: *Mehmet Zeki Erdem, Van Yüzüncü Yıl University, Faculty of Medicine, Department of Otorhinolaryngology, Van, Türkiye* • ***Tel:*** *0532 659 98 91* • ***E-mail:*** *mzekierdem1983@gmail.com* • ***Geliş/Received:*** *02.06.2025* • ***Kabul/Accepted:*** *15.09.2025*

ORCID: *Mehmet Zeki Erdem: 0000-0003-3263-4633 • Abdulaziz Yalınkılıç: 0000-0003-2702-5905 • Yaser Said Çetin: 0000-0002-7684-4600 • Nizamettin Erdem: 0009-0004-7702-642*

Kafkas J Med Sci 2026; 16(1):42–48
doi: 10.5505/kjms.2026.40360

Introduction

Hearing loss is a significant public health issue affecting millions worldwide and significantly reducing their quality of life. According to the World Health Organization (WHO), more than 430 million people, constituting approximately 5% of the world's population, live with hearing loss, and this number is projected to reach 700 million by 2050 (1). Hearing loss can negatively affect individuals' communication skills and social interactions, leading to serious psychosocial and neurological consequences such as social isolation, depression, and cognitive decline (2,3). This situation highlights the critical importance of early diagnosis and effective management of hearing loss in enhancing individuals' quality of life and preventing these negative outcomes.

Hearing loss can be caused by various factors, including genetic factors, aging, infections, exposure to noise, and the use of ototoxic drugs (4,5). Age-related hearing loss (presbycusis) is one of the most common types of hearing loss worldwide, and there is strong scientific evidence that it increases the risk of cognitive decline and dementia in older individuals (2,6). However, it is known that hearing loss affects not only older adults but also younger individuals, such as children and young adults. Hearing loss during childhood can have lasting effects on language development, academic achievement, and social skills, while in young adults, hearing loss can negatively impact workforce productivity and social participation (7). In this context, technological innovations and artificial intelligence (AI)-based tools offer significant potential for the assessment and management of hearing loss.

In recent years, artificial intelligence technologies have led to revolutionary developments in healthcare. In particular, natural language processing, machine learning, and deep learning technologies are widely used in areas such as health data analysis, diagnostic processes, and patient management (8,9). For example, AI-based tools stand out in many areas for their high accuracy, from the analysis of radiological images to the interpretation of genetic data (10). In assessing hearing loss, AI offers significant advantages over traditional methods, providing fast, accurate, and accessible solutions (11). Low-cost, portable AI-based applications have the potential to increase access to healthcare services for individuals with hearing loss (4).

In this study, the effectiveness of three different AI programs used in hearing loss assessment, ChatGPT, Gemini, and Perplexity, has been examined. These programs were compared based on the criteria of quality (GQS), accuracy (Likert), and readability (GFI).

ChatGPT, with its natural language processing capabilities, has broad potential in the healthcare field, while other AI tools like Gemini and Perplexity also stand out for their distinct features. However, the literature on the effectiveness and reliability of these tools for hearing loss assessment is limited. Therefore, this research aims to fill the knowledge gap in this field by comparing the performance of AI-based tools in hearing loss assessment.

Materials and Methods

The 13 most frequently asked questions about hearing loss, particularly on Google, have been identified and directed to AI-based language models (LLMs) such as ChatGPT, Gemini, and Perplexity (Table 1). In the study, only free and publicly accessible versions were preferred. To ensure consistency, the first response given by each language model to each question has been evaluated. The questions were directed to the language models on the same day and through a single user account. The quality and reliability of the obtained responses were evaluated by comparing them with the existing literature by two ear, nose, and throat specialists with at least 10 years of experience, who rated them between 1 and 5 using the global quality scale (GQS) (12). The accuracy of the responses was rated

Table 1. *Frequently asked questions about hearing loss*

No	Question
1	What is hearing loss?
2	What causes hearing loss in the ear?
3	What do degrees of hearing loss mean?
4	What are the symptoms of hearing loss?
5	How to treat hearing loss?
6	Is hearing loss congenital?
7	Does hearing loss progress?
8	How to protect ear and hearing health?
9	Is hearing loss completely healing?
10	How do I know if my child has hearing loss?
11	How is hearing loss diagnosed in babies?
12	Can newborn babies have a hearing test?
13	Do hearing aids cure hearing loss completely?

Table 2. *Comparison of AI programs GQS, GFI, and Likert scores*

	AI Programs	Median	Range	*p
GQS (Quality)	ChatGPT	5.00	1.00	0.018
	Gemini	5.00	1.00	
	Perplexity	4.00	1.00	
GFI (Readability)	ChatGPT	13.13	10.10	0.058
	Gemini	11.71	4.11	
	Perplexity	13.51	8.86	
Likert (Accuracy)	ChatGPT	5.00	1.00	0.072
	Gemini	4.50	1.00	
	Perplexity	4.00	1.00	

* Significance level between groups according to Kruskal-Wallis H test; a, b: Shows the difference between groups according to Bonferroni Post-Hoc test; GQS: global quality scale; GFI: Gunning-Fog index.

on a Likert scale from 1 to 5 (Table 2) (13). The evaluation process was carried out by reaching a consensus among the experts.

The readability of the texts was analyzed using the Gunning-Fog Index (GFI), a widely used measure in the literature that assesses the level of education required to understand a text (14) (Table 3). This index was automatically calculated by transferring the texts to the http://gunning-fog-index.com/ website. This method has enabled objective evaluation of the accessibility of texts for an average reader.

The adequacy of the sample size (n=39), with 13 participants per group, was assessed using G*Power (version 3.1.9.7). Accordingly, in the F-test experimental design, with an effect size of 0.5 (effect size range) and a Type I error (α) of 0.05, the post-hoc power (test power) was determined to be 80.2%, which is considered statistically sufficient. The normality of continuous variables was assessed using the Shapiro-Wilk and Skewness-Kurtosis tests. For data that did not follow a normal distribution, non-parametric tests were preferred. Descriptive statistics for continuous variables were reported as median, range, count (n), and percentage (%). The comparison of continuous variables between groups was performed using the Kruskal-Wallis H test, followed by the Bonferroni-adjusted post-hoc test to identify differences between groups. Categorical variables and relationships between groups were analyzed using the Chi-square test (Fisher's exact test). Statistical significance was set at $p<0.05$, and all analyses were conducted using IBM Statistical Package for Social Sciences (SPSS) program version 26 for Windows.

This study does not require ethics committee approval because it uses only publicly available data and does not involve human participants. During the research process, actions were taken in accordance with international guidelines and regulations to protect the confidentiality and integrity of the data.

Results

In the analysis of the quality and reliability (GQS) scores for artificial intelligence programs, a statistically significant difference was observed between the groups (p=0.018; Table 2). According to the Bonferroni-adjusted Post-Hoc test, the highest GQS score belongs

Table 3. *Distribution of AI programs' by GFI levels*

		Groups						
		ChatGPT		Gemini		Perplexity		
		N	%	N	%	N	%	*p
GFI-Level	College freshman	5	45.5%	3	27.3%	3	27.3%	0.242
	College junior	0	0.0%	0	0.0%	2	100.0%	
	College senior	0	0.0%	0	0.0%	1	100.0%	
	College sophomore	3	60.0%	0	0.0%	2	40.0%	
	Eighth grade	1	50.0%	0	0.0%	1	50.0%	
	High school freshman	0	0.0%	1	50.0%	1	50.0%	
	High school junior	1	14.3%	4	57.1%	2	28.6%	
	High school senior	3	37.5%	3	37.5%	2	25.0%	
	High school sophomore	0	0.0%	3	100.0%	0	0.0%	
	Postgraduate	1	100.0%	0	0.0%	0	0.0%	

* Significance level according to chi-square (Fisher's exact) test results; GFI: Gunning-Fog index.

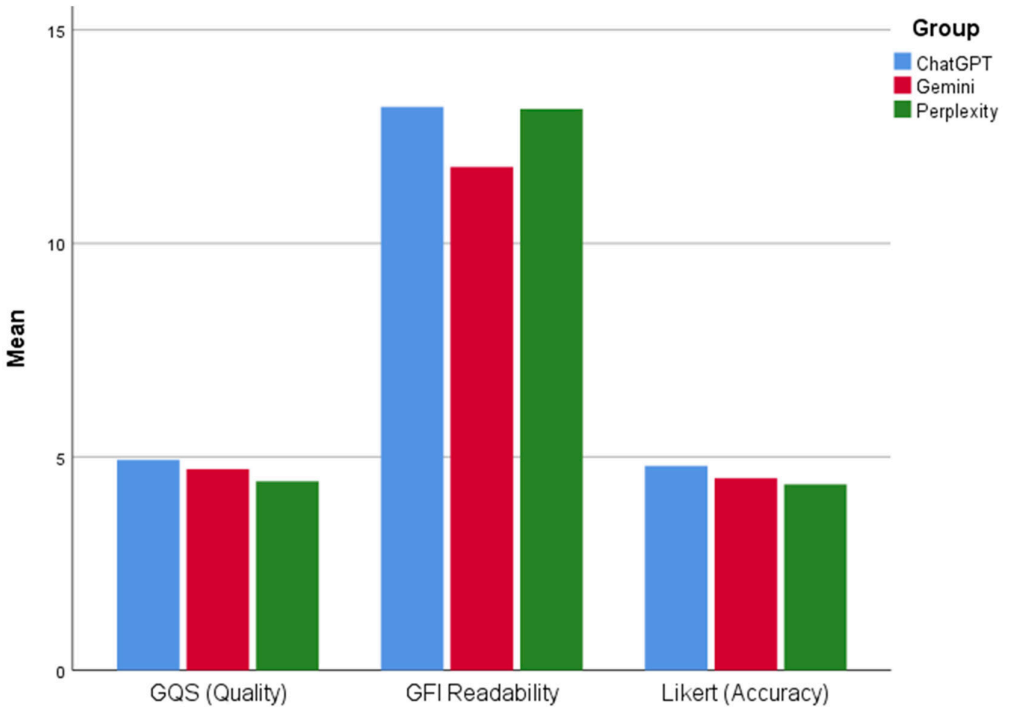

Figure 1. *Comparison of AI programs' GKS, GFI, and Likert scores.*

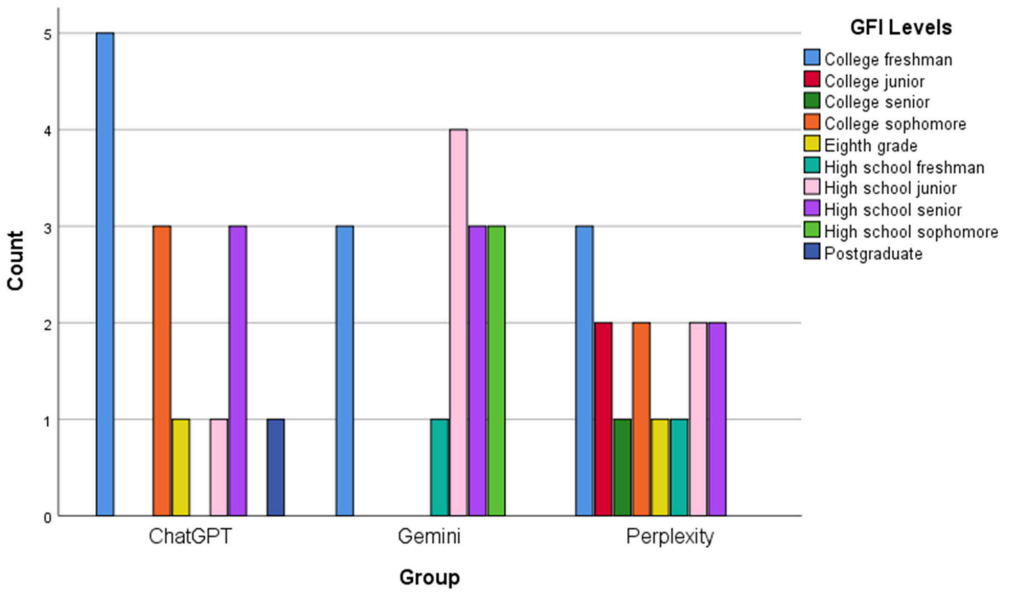

Figure 2. *Graph of AI programs' distribution by GFI Levels.*

to ChatGPT, followed by Gemini and Perplexity. The median value of all groups, 5.00, indicates that artificial intelligence programs generally have a similar level of quality and reliability.

No significant difference was found between the groups in the readability (GFI) measurement (p=0.058, Table 2). This result indicates that artificial intelligence programs exhibit similar readability performance. However, it was observed that Gemini's GFI score was lower than that of other artificial intelligence programs. This may suggest that the texts produced by Gemini perform worse in terms of readability.

Similarly, no significant difference was found between the groups in the accuracy (Likert) measurement

(p=0.072, Table 2). This result indicates that artificial intelligence programs exhibit similar accuracy. However, it was observed that Perplexity's accuracy score was lower than that of other artificial intelligence programs.

No statistically significant difference was found in the distribution of GFI levels across artificial intelligence programs (p=0.242; Table 3). However, it was observed that ChatGPT's texts were more concentrated at the 'University first-year' level, while Gemini's texts were concentrated at the 'High school second year' level. Perplexity's texts, on the other hand, have a more complex structure and show the highest distribution at the 'Postgraduate' level (Table 3, Fig. 1).

Overall, it is understood that ChatGPT performs better in quality and reliability, while Gemini shows lower performance, particularly in readability. It has been observed that Perplexity's texts have a more complex structure and demonstrate lower accuracy than other programs (Fig. 2).

Discussion

The quality assessment was conducted to measure the overall consistency, accuracy, and level of meeting user expectations of the texts produced by artificial intelligence tools. The findings indicate that ChatGPT received the highest quality score (average GKS score: 4.93). This situation may be due to ChatGPT's capacity to comprehend more complex language structures and produce coherent texts with the GPT-4 architecture (15). ChatGPT's superiority suggests it can be an effective tool in areas such as health communication, patient education, and clinical report writing. For example, Moazemi et al. (16) reported that AI-based tools achieve high accuracy in analyzing health data and can be effectively used in clinical decision support systems.

ChatGPT's high-quality score indicates a broader potential for use in the healthcare field. For example, it is thought that this tool could be effectively used for creating patient education materials, writing clinical reports, and providing quick access to information for healthcare professionals (8). However, the lower quality scores of Gemini and Perplexity indicate that these tools have more limited language processing capabilities compared to ChatGPT. However, it should not be forgotten that these tools can be effective in certain contexts. For example, in areas such as producing less complex texts or creating content for a specific user group, these tools may demonstrate adequate performance (17).

In the literature, it has been stated that artificial intelligence tools offer significant advantages in terms of quality in the field of healthcare. For example, Esteva et al. (9) reported that AI-based tools demonstrate performance comparable to that of human experts in areas such as the analysis of radiological images, the interpretation of genetic data, and patient management. However, it should be noted that the performance of these tools depends on the quality and diversity of the datasets used (8). Therefore, testing tools like ChatGPT on larger datasets can enhance their generalizability.

The readability assessment was conducted to measure the comprehensibility and user-friendliness of texts produced by artificial intelligence tools. The findings indicate that ChatGPT and Perplexity perform similarly in terms of readability, but Gemini has a slightly lower score. However, this difference was not statistically significant. These results indicate that AI tools can meet similar readability standards in text production.

Similar readability results indicate that AI tools meet a certain standard in text production. In the literature, it has been noted that AI tools generally demonstrate performance comparable to that of human experts in terms of readability (18). However, it is believed that these tools may be equally effective for different user groups. For example, Perplexity's performance on readability, which is close to ChatGPT despite its lower quality score, suggests that this tool could be used in areas such as creating educational materials. Additionally, similar readability results suggest that these tools may be suitable for individuals across different age groups and educational levels (2).

Similar readability results indicate that more research is needed on how these tools perform across different languages and cultural contexts. In particular, the effectiveness of these tools for content production for multilingual user groups should be evaluated (4). Additionally, more research is needed on how artificial intelligence tools can be used to create educational materials, provide patient information, and improve health literacy (19,20).

The accuracy assessment was conducted to measure the accuracy and reliability of the information produced by artificial intelligence tools. The findings indicate that ChatGPT performed slightly better in accuracy than

other tools, but the difference was not statistically significant. ChatGPT's accuracy score was 4.79, Gemini's 4.50, and Perplexity's 4.36. These results indicate that AI tools exhibit similar accuracy and can be reliably used in healthcare. However, the lack of a significant difference in accuracy suggests that these tools may be equally effective in certain scenarios. In the literature, it has been stated that the accuracy performance of artificial intelligence tools depends on the quality and diversity of the datasets used (8,9). Therefore, studies using broader, more diverse datasets are necessary to improve the accuracy of artificial intelligence tools. Similar accuracy results indicate that these tools can be reliably used in the healthcare field. However, these tools need further accuracy testing before being used in clinical applications. Studies evaluating the effectiveness of these tools for specific health issues, such as hearing loss, are limited (21).

Limitations of the Study

In this study, only the performance of three AI programs (ChatGPT, Gemini, and Perplexity) has been evaluated. Other AI-based tools have not been included in this study. In addition, because the scope of the study's questions is limited to hearing loss, further research is needed to evaluate the performance of AI tools in other health conditions.

Future studies should evaluate artificial intelligence tools using datasets that span different age groups, cultural contexts, and linguistic diversity. Additionally, the ethical and legal dimensions of artificial intelligence tools should also be taken into account. Especially, the privacy and security of health data are significant concerns in AI applications.

Conclusion

This study reveals the potential of artificial intelligence tools for assessing hearing loss. While ChatGPT demonstrates superior quality, other tools also offer similar levels of accuracy and readability. These findings indicate that AI tools can be effectively used in specific health issues, such as hearing loss assessment. However, further research is needed for these tools to be widely used in clinical applications.

References

1. World Health Organization. Deafness and hearing loss. Geneva: World Health Organization; 2023 [Accessed: 17.05.2025]. https://www.who.int
2. Livingston G, Huntley J, Sommerlad A, Ames D, Ballard C, Banerjee S, et al. Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. Lancet. 2020;396(10248):413–446. https://doi.org/10.1016/S0140-6736(20)30367-6
3. Ciorba A, Bianchini C, Pelucchi S, Pastore A. The impact of hearing loss on the quality of life of elderly adults. Clin Interv Aging. 2012;7:159–163. https://doi.org/10.2147/CIA.S26059
4. Wilson BS, Tucci DL, Merson MH, O'Donoghue GM. Global hearing health care: New findings and perspectives. Lancet. 2017;390(10111):2503–2515. https://doi.org/10.1016/S0140-6736(17)31073-5
5. Vos T, Lim SS, Abbafati C, Abbas KM, Abbasi M, Abbasifard M, et al. Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet. 2021;396(10258):1204–1222. https://doi.org/10.1016/S0140-6736(20)30925-9
6. Lin FR, Metter EJ, O'Brien RJ, Resnick SM, Zonderman AB, Ferrucci L. Hearing loss and incident dementia. Arch Neurol. 2011;68(2):214–220. https://doi.org/10.1001/archneurol.2010.362
7. Olusanya BO, Davis AC, Hoffman HJ. Hearing loss: Rising prevalence and impact. Bull World Health Organ. 2019;97(10):646–646A. https://doi.org/10.2471/BLT.19.224683
8. Topol EJ. High-performance medicine: The convergence of human and artificial intelligence. Nat Med. 2019;25(1):44–56. https://doi.org/10.1038/s41591-018-0300-7
9. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nat Med. 2019;25(1):24–29. https://doi.org/10.1038/s41591-018-0316-z
10. Lundervold AS, Lundervold A. An overview of deep learning in medical imaging focusing on MRI. Z Med Phys. 2019;29(2):102–127. https://doi.org/10.1016/j.zemedi.2018.11.002
11. Seo HW, Oh YJ, Oh J, Lee DK, Lee SH, Chung JH, et al. Prediction of hearing recovery with deep learning algorithm in sudden sensorineural hearing loss. Sci Rep. 2024;14(1):20058. https://doi.org/10.1038/s41598-024-70436-0
12. Gudapati JD, Franco AJ, Tamang S, Mikhael A, Hadi MA, Roy V, et al. A study of global quality scale and reliability scores for chest pain: an Instagram-post analysis. Cureus. 2023;15(9):e45629. https://doi.org/10.7759/cureus.45629
13. Jebb AT, Ng V, Tay L. A review of key Likert scale development advances: 1995-2019. Front Psychol. 2021;12:637547. https://doi.org/10.3389/fpsyg.2021.637547

14. Świeczkowski D, Kułacz S. The use of the Gunning Fog Index to evaluate the readability of Polish and English drug leaflets in the context of Health Literacy challenges in Medical Linguistics: an exploratory study. Cardiol J. 2021;28(4):627–631. https://doi.org/10.5603/CJ.a2020.0142
15. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. Adv Neural Inf Process Syst. 2020;33:1877–1901. https://doi.org/10.48550/arXiv.2005.14165
16. Moazemi S, Vahdati S, Li J, Kalkhoff S, Castano LJ, Dewitz B, et al. Artificial intelligence for clinical decision support for monitoring patients in cardiovascular ICUs: A systematic review. Front Med (Lausanne). 2023;10:1109411. https://doi.org/10.3389/fmed.2023.1109411
17. Rashid MM, Atilgan N, Dobres J, Day S, Penkova V, Küçük M, et al. Humanizing AI in education: A readability comparison of LLM and human-created educational content. Proc Hum Factors Ergon Soc Annu Meet. 2024;68(1):596–603. https://doi.org/10.1177/10711813241261689
18. Herbold S, Hautli-Janisz A, Heuer U, Kikteva Z, Trautsch A. A large-scale comparison of human-written versus ChatGPT-generated essays. Sci Rep. 2023;13(1):18617. https://doi.org/10.1038/s41598-023-45644-9
19. Khaja H. Using AI language models to simplify patient education materials. Rheumatology Advisor; 2023. [Accessed: 17.05.2025] https://www.rheumatologyadvisor.com/features/use-of-ai-to-create-patient-education-materials/
20. Moons P, Van Bulck L. Using ChatGPT and Google Bard to improve the readability of written patient information: a proof of concept. Eur J Cardiovasc Nurs. 2024;23(2):122–126. https://doi.org/10.1093/eurjcn/zvad087
21. Maleki Varnosfaderani S, Forouzanfar M. The role of AI in hospitals and clinics: Transforming healthcare in the 21st century. Bioengineering (Basel). 2024;11(4):337. https://doi.org/10.3390/bioengineering11040337