Acta Medica Young Doctors

Orginal Article

Utku Dönger[1], Ahmet Can Doğan[2]

1. Department of Pediatrics, Hatay Mustafa Kemal University Faculty of Medicine, Hatay, Turkey
2. Department of Community Health, Public Health Services, Hatay Provincial Health Directorate

Corresponding Authors: Dr. Utku Dönger, Adress: Hatay Mustafa Kemal University Faculty of Medicine, Alahan, Hatay,

e-Mail: utkudonger@gmail.com

Accepted: 2025-10-25

Dönger U, Doğan AC. Readability Assessment of Large Language Model Responses to Common Postnatal Questions Among Migrant Parents: A Comparative Analysis of ChatGPT and Gemini. Acta Medica Young Dr. 2025;1(3):59-65.

https://doi.org/10.5281/zenodo.17670373

Readability Assessment of Large Language Model Responses to Common Postnatal Questions Among Migrant Parents: A Comparative Analysis of ChatGPT and Gemini

ABSTRACT

Language and communication barriers in migrant communities can substantially hinder access to health information, particularly regarding frequently asked questions in the neonatal period. Large language models (LLMs) have emerged as potential tools to address this information gap; however, the readability of their outputs is critically important from a public health perspective. This study comparatively evaluated the readability of responses generated by ChatGPT and Google Gemini to ten core postnatal questions commonly posed by migrant parents to pediatricians. The ten questions were translated into English and repeatedly submitted to both models on different days and in varying orders. The resulting texts were analyzed using multiple readability tests, including the Automated Readability Index, Flesch Reading Ease, Flesch–Kincaid Grade Level, SMOG Index, and Linsear Write Formula. The findings indicate that Gemini systematically produced more complex texts requiring a higher literacy level in most indices. In contrast, ChatGPT yielded more easily understandable content on scales closely related to health literacy, such as Flesch Reading Ease and SMOG. Overall, both models show potential to support access to health information among migrant populations; however, the complexity of their outputs remains above recommended public health literacy levels, highlighting the need for further optimization and text

simplification, particularly for groups with low health literacy.

Keywords: Migrant health; Readability; Large language models; Postnatal care

Introduction

Armed conflicts and disasters have led to significant population movements, and the health problems of migrants have emerged as an important public health concern. The phenomenon of migration is driven by multiple factors, including political, economic, religious, and social determinants. Since 2009, migration has been recognized as a key social determinant of health (1). In Türkiye, the model of "migrant health workers providing services to their own communities" has been adopted as a distinctive health service approach that substantially reduces communication barriers arising from language and cultural differences. Studies have demonstrated that migrant communities receiving care from health professionals who are themselves migrants at migrant health centers exhibit significantly higher levels of satisfaction with the services provided (2).

In contemporary practice, the internet is often used to find solutions to a wide range of problems quickly. In particular, large language models (LLMs) are increasingly used to obtain quick answers to health-related questions. The quality and readability of these responses are crucial (3). Readability is a construct that can be defined using statistical and psychometric criteria and evaluated through various standardized tests. For example, the Automatic Readability Index is used to measure text comprehensibility; the Flesch Reading Ease Index to quantify difficulty; the Fog Scale to estimate the level of education required for understanding; the Flesch–Kincaid Grade Level to indicate the educational grade necessary for comprehension; the Coleman–Liau Index to assess structural features based on word and character properties; and the SMOG Index to estimate the years of education needed to understand a text (4–9).

After childbirth, parents often experience heightened anxiety about their infants. Asking seemingly simple questions—such as "When can I bathe my baby?" or "How often should I breastfeed?"—may become challenging in the presence of language and communication barriers. This study aimed to analyze the readability of LLM-generated responses to frequently asked postnatal questions in a migrant population in which language and communication barriers are likely to be present.

Methods

This study was designed to repeatedly evaluate how two large language models (LLMs)—ChatGPT, developed by OpenAI, and Gemini, developed by Google—respond to ten core questions frequently asked by migrant parents to pediatricians (10,11). The questions were initially formulated in Turkish based on real-world consultations and then translated into simple, first-person English from the perspective of a parent.

Using a web browser with all cookies and history cleared before each session, the full set of ten questions was submitted to both models on ten different days. On each day, the order of the questions was randomized to minimize potential ordering effects. All responses generated by the models were recorded and subsequently exported into text format for readability analysis.

The ten core questions used in this study were as follows (phrased in first-person language as originally asked by migrant parents):

1. *When can I bathe my baby after birth?*
2. *Should I give my baby any vitamin supplements after birth?*
3. *If my baby and I are both healthy after delivery, when can we be discharged from the hospital?*
4. *If my baby has colic or gas problems, can I give gas drops?*

5. *How can I obtain my baby's birth certificate or birth registration document?*
6. *Has my baby received the first vaccines? How can I obtain the vaccination card? How should I arrange future vaccinations?*
7. *How often should I breastfeed my baby?*
8. *Can I have the newborn hearing test done before we leave the hospital?*
9. *Does my baby have jaundice? How can I know?*
10. *Should I feed my newborn baby only with breast milk? Can I give sugar water? Can I give cow's milk or goat's milk?*

All collected responses for each model (ChatGPT and Gemini) across the ten time points were compiled and prepared for subsequent readability analysis.

Statistical Analysis

All statistical analyses were conducted using SPSS for Windows, version 27.0 (IBM Corp., Armonk, NY, USA). Data were classified as categorical or numerical. Categorical variables were summarized as frequencies and percentages. Numerical variables were examined for distributional characteristics. Variables following a normal distribution were expressed as mean ± standard deviation (SD), whereas non-normally distributed variables were reported as median and interquartile range (IQR).

Comparisons between categorical variables were performed using the chi-square test where appropriate. For repeated measurements and paired comparisons, the paired t-test was applied. A p-value of <0.05 was considered statistically significant.

Results

A total of 200 responses (100 from each model; 10 questions × 10 days) were generated by ChatGPT and Gemini and subjected to readability analysis. The findings revealed statistically significant differences between the two models across several indices.

For the Automated Readability Index (ARI), the mean score for ChatGPT responses was 7.57 ± 1.88, whereas Gemini yielded a mean of 9.88 ± 1.90, and this difference was statistically significant ($p < 0.001$). This indicates that Gemini produced texts requiring a higher reading grade level. Similarly, Flesch–Kincaid Grade Level scores were 7.19 ± 2.12 for ChatGPT and 8.78 ± 1.56 for Gemini, with a statistically significant difference between the two models ($p = 0.005$).

According to the Average Reading Level Consensus, ChatGPT-generated texts have a mean grade level of 8.77 ± 1.25. In contrast, Gemini's mean was 9.94 ± 2.48 ($p = 0.034$), again suggesting that Gemini's outputs required a higher level of literacy. The SMOG Index results showed a similar pattern, with Gemini responses being more complex overall (Gemini: 8.65 ± 1.78 vs. ChatGPT: 7.33 ± 2.34; $p = 0.026$).

In contrast, Flesch Reading Ease scores were significantly higher for ChatGPT (64.00 ± 1.65) than for Gemini (60.00 ± 4.76; $p < 0.001$). Higher scores on this scale correspond to easier-to-read texts; thus, ChatGPT tended to produce relatively more accessible content.

The Linsear Write Formula also indicated a marked difference, with ChatGPT presenting a mean score of 5.49 ± 2.25 compared with 8.03 ± 5.40 for Gemini ($p = 0.03$), again implying greater complexity in Gemini's outputs. For the Coleman–Liau Index, the difference between models did not reach statistical significance (ChatGPT: 9.61 ± 2.25 vs. Gemini: 10.53 ± 1.37; $p = 0.063$), although the scores numerically favored Gemini.

In the Forcast Readability Formula, no significant difference was observed between the two models (ChatGPT: 11.69 ± 1.12 vs. Gemini: 11.56 ± 2.04; $p = 0.402$). As this index is susceptible to word lists and

specialized terminology, the similarity suggests that both models performed comparably when evaluated primarily on vocabulary-based criteria.

Taken together, these results indicate that Gemini generally produced texts at a higher educational grade level and with more complex syntactic structures. In contrast, ChatGPT generated more accessible content, particularly in terms of readability measures closely related to health literacy. These findings underscore the importance of tailoring simplification when utilizing LLM outputs for migrant populations with potential language barriers.

Discussion

The readability of AI-generated health information—particularly the outputs of ChatGPT and Google Gemini—has become a significant focus of recent research in various medical domains. In many clinical contexts, both models have been shown to generate texts that are challenging to read and often require at least college-level literacy, thereby limiting accessibility for the general public. For example, in studies on refractive surgery, both chatbots produced responses that were difficult to read, with no significant difference in overall readability indices. However, ChatGPT tended to provide more detailed explanations, while Gemini offered more concise outputs (12). In the context of emergency medical conditions, both tools produced brochures with comparable readability; however, Gemini's texts were perceived as slightly more straightforward to understand due to a more streamlined structure (13). In frequently asked questions related to refractive surgery, Gemini demonstrated relatively better readability; nevertheless, the readability levels of all models remained higher than recommended for the target patient populations (14).

Similar observations have been made in other fields. In studies examining information for owners of affected dogs, readability scores were comparable across different chatbots; however, the texts were generally difficult to understand for readers with a university reading level or below (15). In the context of ankylosing spondylitis, AI-generated answers exceeded the commonly recommended sixth-grade reading level, emphasizing the need for simplification of outputs to improve public access to health information (16). For retinal detachment, Gemini produced responses that were easier to understand than those of ChatGPT; however, both models still required a relatively high level of education for complete comprehension (17). In breast imaging, Gemini and Microsoft Copilot obtained higher readability scores compared with ChatGPT, suggesting that these models may be more accessible to a broader range of patients (19).

In the context of COVID-19, international models such as ChatGPT and Gemini have been reported to provide more stable content and a more formal documentation style. In contrast, local models were found to be more suitable for public reading due to their more fluent and conversational language (20). Collectively, the literature describes numerous comparative studies across diverse clinical areas, consistently pointing to the tension between the richness of AI-generated information and its readability (3,21).

In this context, the present study adds a novel perspective by focusing specifically on common postnatal questions among migrant parents—a group particularly vulnerable to both health and communication barriers. Our findings demonstrate that Gemini tends to generate texts that require higher levels of literacy, as indicated by several readability indices, including Average Reading Level Consensus, Automated Readability Index, and Flesch–Kincaid Grade Level. By contrast, ChatGPT achieved higher scores on the Flesch Reading Ease test and produced lower (thus more favorable) scores

on scales directly related to health literacy, including the SMOG Index and Linsear Write Formula, indicating relatively more accessible outputs.

From a clinical communication standpoint, particularly in pediatric care and among sensitive groups such as migrant parents, texts intended for patient education should be clear, concise, and tailored to low to moderate literacy levels. The tendency of ChatGPT to generate more readable content represents a practical advantage in supporting communication and shared decision-making in neonatal and pediatric settings. Nonetheless, both models require further refinement to produce texts that align with recommended health literacy standards consistently.

Conclusion

This study comparatively evaluated the readability of responses generated by two large language models—ChatGPT and Google Gemini—to frequently asked postnatal questions from migrant parents. The findings suggest that both models have considerable potential as tools for health communication, yet the overall complexity of their outputs exceeds the recommended readability levels for public health materials. In migrant communities where language barriers are prominent, this may limit access to health services and accurate information.

Readability analyses demonstrated that Gemini produced higher readability scores on many indices, corresponding to a greater literacy requirement. In contrast, ChatGPT yielded lower grade levels and higher Flesch Reading Ease scores, indicating more understandable texts. These results support the need for text simplification, particularly in pediatric care and in communication with vulnerable populations. Consistent with previous studies, our findings underscore that while large language models can provide clinically relevant information, their textual complexity remains a significant barrier for broad public use.

Overall, LLMs offer innovative opportunities to enhance access to health information in migrant health settings. However, improving the readability of AI-generated content remains a clear priority. Future work should focus on developing strategies and tools that promote simplified, culturally adapted, and health-literacy–appropriate outputs, thereby facilitating migrant parents' engagement in neonatal and pediatric care with accessible and trustworthy information.

Table 1. Comparison of comprehensibility (readability) tests

Readability metric	ChatGPT (n = 20) Mean ± SD	Gemini (n = 20) Mean ± SD	p-value
Average Reading Level Consensus	8.77 ± 1.25	9.94 ± 2.48	0.034
Automated Readability Index	7.57 ± 1.88	9.88 ± 1.90	<0.001
Flesch Reading Ease	64.00 ± 1.65	60.00 ± 4.76	<0.001
Fog Scale	9.30 ± 1.98	10.10 ± 1.51	0.079
Flesch–Kincaid Grade Level	7.19 ± 2.12	8.78 ± 1.56	0.005
Coleman–Liau Index	9.61 ± 2.25	10.53 ± 1.37	0.063
SMOG Index	7.33 ± 2.34	8.65 ± 1.78	0.026
Linsear Write Formula	5.49 ± 2.25	8.03 ± 5.40	0.030
Forcast Readability Formula	11.69 ± 1.12	11.56 ± 2.04	0.402

REFERENCES

1. **International Organization for Migration.** *Glossary on Migration.* (trans. ed. Çiçekli B). 2009;22–36. Available from: http://www.goc.gov.tr/files/files/goc_terimleri_sozlugu.pdf [in Turkish]
2. **Keleşmehmet H.** Utilization of health services provided at migrant health centers by Syrian migrants. Specialty thesis. Marmara University Faculty of Medicine, Department of Family Medicine; 2018. Istanbul, Türkiye. [in Turkish]
3. **Aydemir AM.** Warfarin Use: A Readability Comparison of Gemini and ChatGPT. *Acta Medica Young Doctors.* 2025;1(2):59–65. https://doi.org/10.5281/zenodo.17156371
4. **Smith EA, Senter RJ.** Automated readability index. *AMRL TR.* 1967 May:1–14. PMID: 5302480.
5. **McClure G.** Readability formulas: useful or useless (an interview with J. Peter Kincaid). *IEEE Transactions on Professional Communication.* 1987;30:12–15.
6. **Seely J.** Chapter 10: Audience. In: *Oxford Guide to Effective Writing and Speaking: How to Communicate Clearly.* Oxford: Oxford University Press; 2013. p.120–123. ISBN 978-0-19-965270-9.
7. **Kincaid JP, Braby R, Mears J.** Electronic authoring and delivery of technical information. *Journal of Instructional Development.* 1988;11(2):8–13. doi:10.1007/bf02904998.
8. **Coleman M, Liau TL.** A computer readability formula designed for machine scoring. *Journal of Applied Psychology.* 1975;60:283–284.
9. **Ley P, Florio T.** The use of readability formulas in health care. *Psychology, Health & Medicine.* 1996;1(1):7–28. doi:10.1080/13548509608400003.
10. **OpenAI.** ChatGPT (GPT-5.1) [large language model]. Available from: https://www.openai.com
11. **Google DeepMind.** Gemini [large language model]. Available from: https://deepmind.google
12. **Arslan S, Küçükbezirci GU.** Evaluating the accuracy, completeness, and readability of chatbot responses to refractive surgery-related patient questions: a comparative analysis of ChatGPT and Google Gemini. *Cureus.* 2025. https://doi.org/10.7759/cureus.88980
13. **Adithya S, Aggarwal S, Sridhar J, VS K, John V, Singh C.** An observational study to evaluate readability and reliability of AI-generated brochures for emergency medical conditions. *Cureus.* 2024 Aug 31.
14. **Aydın FO, Aksoy BK, Ceylan A, Akbaş YB, Ermiş S, Yıldız BK, et al.** Readability and appropriateness of responses generated by ChatGPT 3.5, ChatGPT 4.0, Gemini, and Microsoft Copilot for FAQs in refractive surgery. *Turkish Journal of Ophthalmology.* 2025;54:313–317.
15. **Erkan Acar EG, Avan BA.** Evaluation of the responses from different chatbots to frequently asked patient questions about impacted canines. *Australasian Orthodontic Journal.* 2025;41(1):288–300. https://doi.org/10.2478/aoj-2025-0020
16. **Kara M, Özduran E, Mercan Kara M, Özbek İC, Hancı V.** Evaluating the readability, quality, and reliability of responses generated by ChatGPT, Gemini, and Perplexity on the most commonly asked questions about ankylosing spondylitis. *PLOS ONE.* 2025;20(6):e0326351. https://doi.org/10.1371/journal.pone.0326351
17. **Strzalkowski P, Strzalkowska A, Chhablani J, Heß K, Errera M-H, Roth M, et al.** Evaluation of the accuracy and readability of ChatGPT-4 and Google Gemini in providing information on retinal detachment: a multicenter expert comparative study. *International Journal of Retina and Vitreous.* 2024;10(1). https://doi.org/10.1186/s40942-024-00579-9

18. **Özduran E, Hancı V, Erkin Y, Özbek İC, Abdulkerimov V.** Assessing the readability, quality and reliability of responses produced by ChatGPT, Gemini, and Perplexity regarding most frequently asked keywords about low back pain. *PeerJ.* 2025;13:e18847. https://doi.org/10.7717/peerj.18847

19. **Tepe M, Emekli E.** Assessing the responses of large language models (ChatGPT-4, Gemini, and Microsoft Copilot) to frequently asked questions in breast imaging: a study on readability and accuracy. *Cureus.* 2024;16. https://doi.org/10.7759/cureus.59960

20. **Liang Z, Kuang Y, Liang X, Ge L, Li Z.** Research on the application of generative artificial intelligence to evaluate responses related to questions about COVID-19 in terms of their accuracy and readability. *Preprints.* 2025. https://doi.org/10.20944/preprints202505.1319.v1

21. **Mutlucan UO.** Evaluation of readability indices of ChatGPT-4 and Google Gemini in cervical disc herniation: evaluation of readability indices of ChatGPT-4 and Google Gemini in cervical disc herniation. *Acta Medica Young Doctors.* 2025;1(2):66–72. https://doi.org/10.5281/zenodo.17156411