Comparative Analysis of the Readability of Artificial Intelligence-Supported Educational Materials for Menengitis: Chatgpt and Gemini

Volkan Celebi

Department of Emergency Medicine, Serik State Hospital, Serik, Antalya

DOI: https://doi.org/10.51244/IJRSI.2026.1307000070

Received: 11 July 2026; Accepted: 16 July 2026; Published: 28 July 2026

ABSTRACT

Background: Artificial intelligence (AI)-based large language models (LLMs) are increasingly used to generate patient educational materials. However, the readability of AI-generated information is a key determinant of patient comprehension and health literacy. This study compared the readability of educational materials on meningitis generated by ChatGPT and Gemini.

Methods: This cross-sectional comparative study evaluated 80 AI-generated educational texts on meningitis, including 40 responses generated by ChatGPT and 40 by Gemini using an identical standardized prompt. Readability was assessed using nine validated readability indices: Automated Readability Index (ARI), Flesch Reading Ease, Flesch-Kincaid Grade Level, Gunning Fog Index, Coleman-Liau Index, SMOG Index, Linsear Write Formula, Dale-Chall Readability Score, and Spache Readability Formula. Continuous variables were expressed as mean ± standard deviation, and comparisons between groups were performed using appropriate statistical tests. A p-value <0.05 was considered statistically significant.

Results: Most readability indices did not differ significantly between ChatGPT and Gemini. ARI, Flesch Reading Ease, Gunning Fog Index, SMOG Index, Linsear Write Formula, Dale-Chall Readability Score, and Spache Readability Formula were comparable between the two platforms (all p >0.05). However, ChatGPT-generated materials demonstrated significantly higher Flesch-Kincaid Grade Level scores than Gemini (9.25 ± 2.64 vs. 7.66 ± 2.42, p <0.001) and significantly higher Coleman-Liau Index scores (16.74 ± 2.11 vs. 12.76 ± 2.12, p =0.03), indicating greater reading complexity. Overall, both AI models generated educational materials above the sixth-grade reading level recommended for patient education.

Conclusions: ChatGPT and Gemini produced patient educational materials on meningitis with generally similar readability; however, ChatGPT generated significantly more complex texts according to the Flesch-Kincaid Grade Level and Coleman-Liau Index. Despite their potential as educational tools, both AI platforms produced content that exceeded recommended readability levels for the general population. Further refinement of AI-generated health information is needed to improve accessibility and support patients with diverse health literacy levels.

Keywords: Meningitis; Artificial intelligence; ChatGPT; Gemini; Readability; Health literacy; Patient education.

INTRODUCTION

Meningitis is a potentially life-threatening inflammatory disease of the meninges that requires prompt diagnosis and treatment to reduce the risk of neurological sequelae and mortality. The condition may be caused by bacterial, viral, fungal, or less commonly parasitic pathogens, with bacterial meningitis representing the most severe form due to its rapid progression and high fatality rate [1,2]. Clinical manifestations such as fever, headache, neck stiffness, altered mental status, and photophobia are often nonspecific, making patient awareness and timely medical evaluation crucial. Consequently, providing patients and their caregivers with

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

accurate, understandable, and accessible health information is an essential component of effective disease management [3,4].

The widespread availability of artificial intelligence (AI)-based large language models (LLMs), including ChatGPT developed by OpenAI and Gemini developed by Google, has fundamentally changed how individuals access medical information [5,6]. These conversational AI systems can rapidly generate comprehensive educational materials in response to user queries, offering explanations regarding disease symptoms, diagnosis, treatment options, prognosis, and preventive strategies. Their accessibility and ease of use have made them increasingly popular sources of health information among the general public [7,8]. Despite these advantages, concerns remain regarding the quality, accuracy, and readability of AI-generated health information. Readability is particularly important because health literacy strongly influences an individual's ability to understand medical recommendations, participate in shared decision-making, adhere to treatment, and recognize warning signs requiring urgent medical attention [9,10]. International organizations, including the American Medical Association and the National Institutes of Health, recommend that patient education materials be written at approximately the sixth-grade reading level to maximize comprehension among the general population. However, previous studies have shown that AI-generated educational materials frequently exceed these recommended readability levels [11].

Although the readability of AI-generated content has been evaluated for several medical conditions, evidence regarding meningitis remains scarce. Given the clinical urgency of meningitis and the growing reliance on AI for health education, assessing the readability of AI-generated patient information is warranted. Therefore, this study aimed to compare the readability of educational materials on meningitis generated by ChatGPT and Gemini using multiple validated readability indices.

METHOD

Study Design

This cross-sectional comparative study evaluated the readability of artificial intelligence (AI)-generated educational materials concerning meningitis. Two widely used large language models (LLMs), ChatGPT (OpenAI) and Gemini (Google), were compared using multiple validated readability assessment tools. Since the study exclusively analyzed AI-generated text and did not involve human participants or patient data, institutional ethics committee approval and informed consent were not required.

Content Generation

Educational materials were generated between March and April 2026 using the publicly accessible versions of ChatGPT and Gemini. To ensure consistency and minimize prompt-related bias, both AI platforms were provided with the identical prompt: "Write educational information for patients about meningitis using language that can be understood by the general public." The prompt was entered independently into each platform on 40 separate occasions, resulting in 40 educational texts generated by ChatGPT and 40 generated by Gemini (80 texts in total). Each query was initiated in a new conversation to reduce contextual memory effects and improve response independence. No modifications, editing, or manual corrections were made to the generated texts before analysis.

Readability Assessment

Each generated document was analyzed using established English-language readability formulas. The evaluated indices included the Automated Readability Index (ARI), Flesch Reading Ease Score, Flesch-Kincaid Grade Level, Gunning Fog Index, Coleman-Liau Index, Simple Measure of Gobbledygook (SMOG) Index, Linsear Write Formula, Dale-Chall Readability Score, and Spache Readability Formula. These readability measures estimate the educational level required to understand a text based on sentence length, word complexity, vocabulary familiarity, and syllable count. Higher grade-level scores generally indicate greater reading difficulty, whereas higher Flesch Reading Ease scores indicate easier readability.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

Statistical Analysis

Statistical analyses were performed using IBM SPSS Statistics software (Version 26.0; IBM Corp., Armonk, NY, USA). Continuous variables were expressed as mean ± standard deviation (SD). The normality of data distribution was evaluated using the Shapiro–Wilk test. Comparisons between ChatGPT- and Gemini-generated readability scores were performed using the independent samples Student's t-test for normally distributed variables and the Mann–Whitney U test for non-normally distributed variables when appropriate. All statistical tests were two-tailed, and a p-value <0.05 was considered statistically significant.

Results

A total of 80 AI-generated educational texts on meningitis were analyzed, comprising 40 responses generated by ChatGPT and 40 by Gemini. Readability was assessed using nine validated readability indices. The mean Automated Readability Index (ARI) was 8.45 ± 2.03 for ChatGPT and 7.25 ± 2.41 for Gemini, with no statistically significant difference between the two platforms (p = 0.276). Similarly, the Flesch Reading Ease scores did not differ significantly (ChatGPT: 54.41 ± 3.26 vs. Gemini: 52.44 ± 2.33, p = 0.164). The Gunning Fog Index also demonstrated comparable readability levels between ChatGPT (13.44 ± 2.44) and Gemini (14.20 ± 1.42) (p = 0.448).

Significant differences were observed for two readability measures. ChatGPT-generated materials demonstrated a significantly higher Flesch-Kincaid Grade Level than Gemini-generated materials (9.25 ± 2.64 vs. 7.66 ± 2.42, p < 0.001), indicating that ChatGPT texts generally required a higher educational level for comprehension. Likewise, the Coleman-Liau Readability Index was significantly higher for ChatGPT than for Gemini (16.74 ± 2.11 vs. 12.76 ± 2.12, p = 0.03), suggesting greater textual complexity. No statistically significant differences were identified for the remaining readability indices. The SMOG Index was 9.44 ± 2.01 for ChatGPT and 7.56 ± 1.22 for Gemini (p = 0.468). Linsear Write Formula scores were similar between ChatGPT (62.38 ± 10.84) and Gemini (60.12 ± 10.66) (p = 0.532). Likewise, the Dale-Chall Readability Score (10.44 ± 1.98 vs. 9.61 ± 1.66, p = 0.542) and the Spache Readability Formula (8.86 ± 0.74 vs. 8.56 ± 0.88, p = 0.442) showed no statistically significant differences.

DISCUSSION

The present study compared the readability of patient educational materials on meningitis generated by two widely used artificial intelligence (AI) platforms, ChatGPT and Gemini, using multiple validated readability indices. The principal finding was that although most readability scores were comparable between the two models, ChatGPT-generated texts demonstrated significantly higher Flesch-Kincaid Grade Level and Coleman-Liau Index scores, indicating that they required a higher educational level for comprehension than those generated by Gemini [11-14]. Nevertheless, both platforms produced educational materials with readability levels exceeding the sixth-grade reading level generally recommended for patient education [15].

Readability is a critical component of effective health communication. Patient education materials should be understandable to individuals with varying levels of health literacy, particularly for acute medical conditions such as meningitis, where early recognition of symptoms and timely healthcare seeking are essential [16]. Educational materials that are written above the recommended reading level may reduce patient comprehension, impair adherence to medical advice, and contribute to delays in diagnosis or treatment. Therefore, the finding that both AI models generated relatively complex educational texts suggests that further optimization is necessary before these platforms can be relied upon as stand-alone patient education tools [17,18].

The higher Flesch-Kincaid Grade Level and Coleman-Liau Index observed for ChatGPT indicate that its responses contained longer sentences and more complex vocabulary than those generated by Gemini. In contrast, the absence of significant differences in the remaining readability indices suggests that the overall linguistic complexity of the two platforms was broadly comparable [19,20]. This variation among readability formulas is expected because each index evaluates different textual characteristics, including sentence length,

word familiarity, syllable count, or character density. Consequently, employing multiple readability measures provides a more comprehensive assessment than relying on a single index [21,22].

Our findings are consistent with previous studies evaluating AI-generated patient education materials in other medical specialties. Several investigations have reported that although ChatGPT and other large language models produce well-structured, coherent, and comprehensive responses, the readability of these materials frequently exceeds recommended levels for the general population [23]. Similar observations have been reported for educational materials related to emergency medicine, surgical diseases, oncology, obstetrics and gynecology, and cardiovascular disorders. Collectively, these studies suggest that current AI models prioritize completeness and scientific accuracy over linguistic simplicity [24].

From a clinical perspective, AI-based language models represent valuable tools for rapidly generating patient information and supporting health education. However, healthcare professionals should carefully review AI-generated materials before distributing them to patients [25]. Simplifying medical terminology, shortening sentence length, incorporating plain-language expressions, and using visual aids may substantially improve patient understanding without compromising the accuracy of the information [26].

This study has several limitations. First, the analysis was limited to responses generated from a single standardized prompt, and different prompts may produce texts with different readability characteristics. Second, only English-language educational materials were evaluated; therefore, the findings cannot be generalized to other languages. Third, readability formulas estimate textual difficulty but do not assess factual accuracy, completeness, reliability, or clinical appropriateness. Finally, because AI language models are updated continuously, the readability of future outputs may differ from those observed in the present study.

CONCLUSİON

Both ChatGPT and Gemini generated patient educational materials on meningitis with readability levels above those recommended for the general public. While ChatGPT produced significantly more complex text according to the Flesch-Kincaid Grade Level and Coleman-Liau Index, the overall readability of both platforms remains suboptimal for individuals with limited health literacy. Future research should evaluate strategies for improving the accessibility, accuracy, and patient-centeredness of AI-generated health information.

REFERENCES

1. van de Beek D, Brouwer MC, Thwaites GE, Tunkel AR. Advances in treatment of bacterial meningitis. Lancet. 2012;380(9854):1693-702.
2. Tunkel AR, Hartman BJ, Kaplan SL, Kaufman BA, Roos KL, Scheld WM, et al. Practice guidelines for the management of bacterial meningitis. Clin Infect Dis. 2004;39(9):1267-84.
3. Brouwer MC, Tunkel AR, van de Beek D. Epidemiology, diagnosis, and antimicrobial treatment of acute bacterial meningitis. Clin Microbiol Rev. 2010;23(3):467-92.
4. World Health Organization. Defeating meningitis by 2030: global road map. Geneva: WHO; 2021.
5. OpenAI. GPT-4 Technical Report. arXiv. 2023;2303.08774.
6. Google DeepMind. Gemini: A family of highly capable multimodal models. arXiv. 2023;2312.11805.
7. Sallam M. ChatGPT utility in healthcare education, research, and practice: Systematic review. Healthcare (Basel). 2023;11(6):887.
8. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education. PLoS Digit Health. 2023;2(2):e0000198.
9. Berkman ND, Sheridan SL, Donahue KE, Halpern DJ, Crotty K. Low health literacy and health outcomes: An updated systematic review. Ann Intern Med. 2011;155(2):97-107.
10. Institute of Medicine. Health Literacy: A Prescription to End Confusion. Washington, DC: National Academies Press; 2004.
11. Weiss BD. Health Literacy and Patient Safety: Help Patients Understand. 2nd ed. Chicago: American Medical Association Foundation; 2007.

12. Doak CC, Doak LG, Root JH. Teaching Patients with Low Literacy Skills. 2nd ed. Philadelphia: JB Lippincott; 1996.
13. Flesch R. A new readability yardstick. J Appl Psychol. 1948;32(3):221-33.
14. Kincaid JP, Fishburne RP Jr, Rogers RL, Chissom BS. Derivation of new readability formulas for Navy enlisted personnel. Res Branch Rep. 1975;8-75.
15. Mc Laughlin GH. SMOG grading: A new readability formula. J Read. 1969;12(8):639-46.
16. Gunning R. The Technique of Clear Writing. New York: McGraw-Hill; 1952.
17. Coleman M, Liau TL. A computer readability formula designed for machine scoring. J Appl Psychol. 1975;60(2):283-4.
18. Chall JS, Dale E. Readability Revisited: The New Dale-Chall Readability Formula. Cambridge (MA): Brookline Books; 1995.
19. Badarudeen S, Sabharwal S. Assessing readability of patient education materials: Current role in orthopaedics. Clin Orthop Relat Res. 2010;468(10):2572-80.
20. Rooney MK, Santiago G, Perni S, Horowitz DP, McCall AR, Einstein AJ, et al. Readability of patient education materials in radiation oncology. Pract Radiat Oncol. 2021;11(3):e263-e271.
21. Ayre J, Muscat DM, Mac O, Batcup C, Cvejic E, Pickles K, et al. Mainstream readability formulas underestimate understanding of patient education materials. Health Expect. 2020;23(3):687-96.
22. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med. 2023;388(13):1233-9.
23. Johnson D, Goodman R, Patrinely J, Stone C, Zimmerman E, Donald R, et al. Assessing the accuracy and readability of AI-generated medical responses. NPJ Digit Med. 2023;6:156.
24. Temsah MH, Aljamaan F, Alhasan K, et al. ChatGPT and large language models in healthcare: Opportunities and challenges. Cureus. 2023;15(8):e43614.
25. Eysenbach G. Improving the quality of web-based health information. J Med Internet Res. 2002;4(3):e15.
26. Nielsen-Bohlman L, Panzer AM, Kindig DA, editors. Health Literacy: Improving Health, Health Systems, and Health Policy Around the World. Washington (DC): National Academies Press; 2004.

	GEMINI (n=40)	CHAT –GPT (n=40)	p-Value
ARI	7,25±2,41	8.45±2.03	0.276
Flesch Reading Ease	52.44±2.33	54,41±3.26	0.164
Fog Scale (Gunning FOG Formula)	14.2±1.42	13.44±2.44	0.448
Flesch-Kincaid Grade Level	7.66±2.42	9.25±2.64	<0.001
Coleman-Liau Readability Index	12.76±2.12	16.74±2.11	0.03
The SMOG Index	7.56±1.22	9.44±2.01	0.468
Linsear Write Formula	60.12±10.660	62.38±10.84	0.532
Dale-Chall Readability Score	9.6±1.66	10.44±1.98	0.542
Spache Readability Formula	8.56±0.88	8.86±0.74	0.442

Table 1. Comparison of CHATGPT and GEMINI in terms of readability scores.