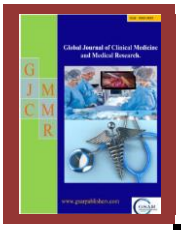

Global Journal of Clinical Medicine and Medical Research [GJCMMR]
ISSN: 2583-987X (Online)
Abbreviated key title: Glob.J.Clinic.Medici.Medica.Res.
Frequency: Monthly
Published By GSAR Publishers
Journal Homepage Link- https://gsarpublishers.com/journal-gjcmmr-home/

Comparative Analysis of the Readability of Artificial Intelligence-Supported Educational Materials for Patients with Tubo-ovaryan Apse: ChatGPT versus Gemini

By

Zeynel Umut Canbaba
Department of Obstetrics and gynaecology, Muş State Hospital, Muş, Türkiye

Article History

Received: 01/07/2026
Accepted: 06/07/2026
Published: 08/07/2026

Vol – 4 Issue – 7

PP: -06-10

Abstract

Background: *Tubo-ovarian abscess (TOA) is a serious gynecological emergency requiring timely diagnosis, appropriate treatment, and effective patient education. Large language models (LLMs), such as ChatGPT and Gemini, are increasingly used to provide medical information; however, the readability of their educational content remains insufficiently investigated.*

Objective: *This study aimed to compare the readability of patient educational materials on tubo-ovarian abscess generated by ChatGPT and Gemini using multiple validated readability indices.*

Methods: *In this cross-sectional comparative study, 50 frequently asked questions regarding tubo-ovarian abscess were submitted independently to ChatGPT and Gemini, resulting in 100 AI-generated educational texts. Responses were analyzed using nine validated readability indices, including the Automated Readability Index (ARI), Flesch Reading Ease Score, 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 readability scores were compared using appropriate statistical tests. A* p*-value <0.05 was considered statistically significant.*

Results: *Most readability indices demonstrated comparable performance between ChatGPT and Gemini. No statistically significant differences were observed in ARI (*p*=0.664), Flesch Reading Ease (*p*=0.364), Gunning Fog Index (*p*=0.468), SMOG Index (*p*=0.464), Linsear Write Formula (*p*=0.112), Dale–Chall Readability Score (*p*=0.148), or Spache Readability Formula (*p*=0.684). However, Gemini generated significantly more readable educational materials according to the Flesch–Kincaid Grade Level (6.46±2.64 vs. 9.66±4.64;* p*<0.001) and Coleman–Liau Index (11.64±3.64 vs. 16.24±2.64;* p*=0.012). Despite these differences, both AI models produced texts requiring reading abilities above the sixth-grade level recommended for patient education.*

Conclusions: *Both ChatGPT and Gemini generated informative educational materials for patients with tubo-ovarian abscess; however, their readability generally exceeded recommended health literacy standards. Gemini demonstrated superior readability in selected grade-level indices, suggesting that it may provide more accessible patient education than ChatGPT. Nevertheless, further refinement of AI-generated medical content is necessary to improve accessibility for individuals with limited health literacy while maintaining clinical accuracy.*

Keywords: *Tubo-ovarian abscess; Artificial intelligence; ChatGPT; Gemini; Readability; Health literacy; Patient education; Large language models.*

Introduction

Tubo-ovarian abscess (TOA) is one of the most severe complications of pelvic inflammatory disease (PID) and represents a significant cause of gynecological morbidity among women of reproductive age [1]. Characterized by the formation of an inflammatory mass involving the fallopian tube, ovary, and adjacent pelvic structures, TOA can result in infertility, chronic pelvic pain, ectopic pregnancy, sepsis, and even death if not diagnosed and treated promptly [2]. Although advances in antimicrobial therapy and minimally invasive surgical techniques have improved clinical outcomes, successful management also depends on patients' understanding of their condition, treatment options, and the importance of adherence to follow-up recommendations [3]. Patient education has become an essential component of modern healthcare, contributing to improved treatment compliance, shared decision-making, and patient satisfaction. Traditionally, educational materials have been developed by

*Corresponding Author: **Zeynel Umut Canbaba**

healthcare professionals or medical organizations; however, the rapid expansion of internet-based health information has transformed the way patients obtain medical knowledge. Increasingly, patients seek online resources before or after clinical consultations to better understand their diagnoses. Unfortunately, the quality, accuracy, and readability of these resources vary considerably, and many exceed the average reading ability of the general population [4,5].

Health literacy, defined as an individual's capacity to obtain, process, and understand basic health information necessary for making appropriate health decisions, is recognized as a major determinant of healthcare outcomes [6]. Several international organizations, including the American Medical Association and the National Institutes of Health, recommend that patient educational materials be written at approximately the sixth-grade reading level to maximize comprehension across diverse populations. Materials that require advanced reading skills may reduce understanding, impair adherence to treatment recommendations, and contribute to poorer clinical outcomes [7,8]. Recent developments in artificial intelligence (AI), particularly large language models (LLMs), have introduced new opportunities for generating patient-centered educational content. Conversational AI systems such as ChatGPT (OpenAI) and Gemini (Google) are capable of producing detailed explanations of complex medical topics within seconds and are increasingly being used by patients seeking immediate health information [9,10]. These systems can generate personalized responses, simplify medical terminology, and provide interactive educational experiences that differ substantially from conventional static online resources. Consequently, AI-generated educational materials may play an increasingly important role in patient education in the near future [11,12].

Comparisons between different AI platforms are particularly valuable because each model employs distinct training methodologies, language-generation strategies, and response optimization techniques, which may influence the complexity and readability of the information they produce [13,14]. Understanding these differences may help clinicians identify AI systems that provide patient education at a more appropriate reading level and may guide future improvements in AI-assisted health communication. Therefore, the present study aimed to compare the readability of educational materials regarding tubo-ovarian abscess generated by two widely used large language models, ChatGPT and Gemini, using multiple validated readability indices [15,16]. By evaluating whether these AI platforms produce information that aligns with recommended health literacy standards, this study seeks to contribute to the growing body of evidence on the role of artificial intelligence in patient education and to provide insights into the optimization of AI-generated healthcare information.

Methods

Study Design

This cross-sectional comparative study evaluated the readability of patient educational materials on tubo-ovarian abscess (TOA) generated by two large language models (LLMs), ChatGPT (OpenAI) and Gemini (Google). The study was conducted between May and June 2026. Since the research involved publicly available artificial intelligence systems and did not include human participants, patient data, or identifiable personal information, institutional ethics committee approval was not required.

Question Development and AI Response Generation

A total of 50 frequently asked questions regarding tubo-ovarian abscess were developed based on patient education resources published by international gynecology organizations, clinical practice guidelines, and common questions encountered in routine gynecological practice. The questions covered disease definition, causes, symptoms, diagnosis, treatment options, antibiotic therapy, surgical management, fertility implications, possible complications, prognosis, prevention, and follow-up recommendations. To ensure consistency, identical prompts were entered separately into ChatGPT and Gemini using newly initiated conversations to minimize contextual memory effects. Each question was submitted only once to each platform, and responses were generated using the default model settings without additional prompting or refinement.

Text Preparation

The generated responses were exported into Microsoft Word documents. Formatting elements that could influence readability calculations, including bullet points, hyperlinks, tables, emojis, and special characters, were removed while preserving the original wording and sentence structure. No grammatical corrections or manual simplifications were performed to maintain the authenticity of the AI-generated content. Readability analyses were performed using validated English-language readability formulas widely employed in health communication research. 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 indices estimate text difficulty based on variables such as sentence length, word length, syllable count, and vocabulary complexity. Lower grade-level scores and higher Flesch Reading Ease values indicate materials that are easier for the general public to understand.

Statistical Analysis

Statistical analyses were performed using IBM SPSS Statistics version 29.0 (IBM Corp., Armonk, NY, USA). Continuous variables were expressed as mean ± standard deviation (SD). The normality of data distribution was assessed using the Shapiro–Wilk test. Comparisons of readability scores between ChatGPT and Gemini were performed using the independent-samples t-test for normally distributed variables or the Mann–Whitney U test when appropriate. A two-sided *p*-value <0.05 was considered statistically significant. The primary outcome was the comparison of readability scores between the two AI platforms across the nine validated readability indices.

*Corresponding Author: **Zeynel Umut Canbaba**

Results

A total of 100 patient educational texts were analyzed, consisting of 50 responses generated by ChatGPT and 50 responses generated by Gemini for identical questions regarding tubo-ovarian abscess. Readability was evaluated using nine validated readability indices, and the comparative results are presented in **Table 1.** The **Automated Readability Index (ARI)** demonstrated no statistically significant difference between the two AI platforms (Gemini: 9.41 ± 2.12 vs. ChatGPT: 9.78 ± 1.24; $p = 0.664$). Similarly, the **Flesch Reading Ease Score,** which indicates greater readability with higher scores, was comparable between Gemini (56.25 ± 2.66) and ChatGPT (54.82 ± 4.66), with no significant difference ($p = 0.364$). The **Gunning Fog Index** also showed similar readability levels (11.22 ± 2.42 vs. 13.24 ± 2.96; $p = 0.468$).

Significant differences were observed in two grade-level readability measures. The **Flesch–Kincaid Grade Level** was significantly lower for Gemini than for ChatGPT (6.46 ± 2.64 vs. 9.66 ± 4.64, respectively; $p < 0.001$), indicating that Gemini-generated educational materials required a lower educational level for comprehension. Likewise, the **Coleman–Liau Readability Index** was significantly lower in Gemini-generated texts (11.64 ± 3.64) than in ChatGPT-generated texts (16.24 ± 2.64; $p = 0.012$), suggesting that Gemini produced less linguistically complex content.No statistically significant differences were identified for the remaining readability metrics. The **SMOG Index** was 8.84 ± 2.12 for Gemini and 9.46 ± 2.56 for ChatGPT ($p = 0.464$). The **Linsear Write Formula** scores were 62.32 ± 10.64 and 64.65 ± 11.84, respectively ($p = 0.112$). Similarly, the **Dale–Chall Readability Score** (11.22 ± 0.88 vs. 11.56 ± 1.02; $p = 0.148$) and the **Spache Readability Formula** (7.96 ± 2.48 vs. 7.56 ± 2.24; $p = 0.684$) did not differ significantly between the two AI models.

Discussion

The present study compared the readability of patient educational materials on tubo-ovarian abscess (TOA) generated by two widely used large language models (LLMs), ChatGPT and Gemini [12,13]. Overall, both platforms produced educational texts with readability levels exceeding the sixth-grade reading level recommended for patient-directed health information. Although most readability indices showed no statistically significant differences between the two models, Gemini generated significantly more readable content according to the Flesch–Kincaid Grade Level and Coleman–Liau Index. These findings suggest that while both AI systems are capable of producing comprehensive educational materials, Gemini may be better suited for communicating complex gynecological information in language that is more accessible to the general public [17].

Health literacy is a fundamental determinant of patient engagement, treatment adherence, and healthcare outcomes. Patients with TOA are often confronted with unfamiliar medical terminology, complex treatment regimens, and concerns regarding fertility preservation, hospitalization, and potential surgical intervention [18]. Consequently, educational materials must present accurate information in language that patients can readily understand. The American Medical Association and the National Institutes of Health recommend that patient education resources be written at approximately the sixth-grade reading level. In the present study, both ChatGPT and Gemini exceeded this recommendation in several readability metrics, indicating that further simplification is needed before AI-generated educational materials can be considered fully appropriate for patients with limited health literacy [19]. Our findings are consistent with previous investigations evaluating AI-generated patient education across various medical specialties. Several studies have demonstrated that ChatGPT is capable of producing medically accurate and well-structured educational content but frequently generates responses requiring reading abilities above those recommended for the general population. Similar observations have been reported in orthopedic surgery, oncology, gastroenterology, urology, ophthalmology, and emergency medicine, where AI-generated materials were considered informative but linguistically demanding. These studies suggest that although LLMs have considerable potential as patient education tools, readability remains a significant limitation that must be addressed before widespread clinical implementation.

The superior performance of Gemini in the Flesch–Kincaid Grade Level and Coleman–Liau Index may reflect differences in the underlying language generation strategies used by the two models [20]. These readability formulas primarily evaluate sentence length, word complexity, and character-based linguistic features. Gemini appeared to produce shorter sentences and employ less complex vocabulary, resulting in lower educational grade requirements. In contrast, ChatGPT tended to generate longer and more detailed explanations that, while potentially increasing informational completeness, also increased linguistic complexity [21]. This observation highlights an important trade-off between comprehensiveness and readability in AI-generated health communication. Interestingly, no statistically significant differences were observed in the Automated Readability Index, Flesch Reading Ease, Gunning Fog Index, SMOG Index, Dale–Chall Readability Score, Linsear Write Formula, or Spache Readability Formula [22]. The consistency across these measures suggests that both AI platforms produce educational materials with broadly similar linguistic characteristics despite differences detected by selected grade-level indices. This also emphasizes the importance of using multiple validated readability formulas rather than relying on a single metric when evaluating patient education materials [23].

The present study has important clinical implications. As patients increasingly seek health information from conversational AI systems before or after medical consultations, clinicians should recognize that AI-generated educational materials may not always be appropriate for individuals with lower literacy levels [24]. Healthcare professionals should therefore review AI-generated content before recommending it to patients or consider modifying the language to improve accessibility. Future AI systems could

*Corresponding Author: **Zeynel Umut Canbaba**

incorporate automated readability optimization algorithms that dynamically adjust vocabulary, sentence structure, and explanation depth according to the user's literacy level while maintaining medical accuracy [25].

This study has several limitations. First, only two LLMs were evaluated, and the findings may not be generalizable to other AI platforms. Second, readability was assessed using established linguistic formulas that estimate reading difficulty but do not directly measure patient comprehension, understanding, or knowledge retention. Third, the study focused exclusively on English-language responses regarding TOA; readability may differ in other languages or across different gynecological conditions. Finally, because AI models are continuously updated, the readability of generated responses may change over time, potentially affecting reproducibility.

Conclusion

Both ChatGPT and Gemini generated patient educational materials on tubo-ovarian abscess that generally exceeded recommended readability levels for patient education. Although Gemini demonstrated significantly better readability according to the Flesch–Kincaid Grade Level and Coleman–Liau Index, neither platform consistently achieved the sixth-grade reading level recommended for optimal patient comprehension. Future improvements in large language models should prioritize readability alongside factual accuracy to enhance their effectiveness as reliable patient education tools in gynecological practice.

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

	GEMINI (n=50)	CHAT –GPT (n=50)	p-Value
ARI	9.41±2.12	9.78±1.24	0.664
Flesch Reading Ease	56.25±2.66	54.82±4.66	0.364
Fog Scale (Gunning FOG Formula)	11.22±2.42	13.24±2.96	0.468
Flesch-Kincaid Grade Level	6.46±2.64	9.66±4.64	<0.001
Coleman-Liau Readability Index	11.64±3.64	16.24±2.64	0.012
The SMOG Index	8.84±2.12	9.46±2.56	0.464
Linsear Write Formula	62.32±10.64	64.65±11.84	0.112
Dale-Chall Readability Score	11.22±0.88	11.56±1.02	0.148
Spache Readability Formula	7.96±2.48	7.56±2.24	0.684

References

1. Haggerty CL, Taylor BD. Mycoplasma genitalium: an emerging cause of pelvic inflammatory disease. **Infect Dis Obstet Gynecol.** 2011;2011:959816.
2. Sweet RL. Pelvic inflammatory disease: current concepts of diagnosis and management. **Curr Infect Dis Rep.** 2012;14(2):194-203.
3. Committee on Practice Bulletins—Gynecology. Pelvic inflammatory disease. **Obstet Gynecol.** 2018;131(6):e174-e180.
4. 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.
5. Nielsen-Bohlman L, Panzer AM, Kindig DA. **Health Literacy: A Prescription to End Confusion.** Washington (DC): National Academies Press; 2004.
6. Weiss BD. Health literacy and patient safety: help patients understand. Manual for clinicians. 2nd ed. Chicago: American Medical Association Foundation; 2007.
7. Doak CC, Doak LG, Root JH. **Teaching Patients with Low Literacy Skills.** 2nd ed. Philadelphia: JB Lippincott; 1996.
8. McInnes N, Haglund BJ. Readability of online health information: implications for health literacy. **Inform Health Soc Care.** 2011;36(4):173-189.
9. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. **Healthcare (Basel).** 2023;11(6):887.
10. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. **PLoS Digit Health.** 2023;2(2):e0000198.
11. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination? **JMIR Med Educ.** 2023;9:e45312.
12. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot

*Corresponding Author: **Zeynel Umut Canbaba**

responses to patient questions posted to a public social media forum. **JAMA Intern Med.** 2023;183(6):589-596.
13. 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):1233-1239.
14. OpenAI. GPT-4 Technical Report. **arXiv.** 2023;arXiv:2303.08774.
15. Google DeepMind. Gemini: A family of highly capable multimodal models. **arXiv.** 2023;arXiv:2312.11805.
16. Flesch R. A new readability yardstick. **J Appl Psychol.** 1948;32(3):221-233.
17. Kincaid JP, Fishburne RP Jr, Rogers RL, Chissom BS. Derivation of new readability formulas for Navy enlisted personnel. **Res Branch Rep.** 1975;8-75.
18. Gunning R. **The Technique of Clear Writing.** New York: McGraw-Hill; 1952.
19. Mc Laughlin GH. SMOG grading: a new readability formula. **J Read.** 1969;12(8):639-646.
20. Coleman M, Liau TL. A computer readability formula designed for machine scoring. **J Appl Psychol.** 1975;60(2):283-284.
21. Chall JS, Dale E. **Readability Revisited: The New Dale-Chall Readability Formula.** Cambridge (MA): Brookline Books; 1995.
22. Spache G. **A New Readability Formula for Primary Grade Reading Materials.** Elem Sch J. 1953;53(7):410-413.
23. AlRyalat SAS, Alhuzaim WM, Alzahrani AJ, et al. Readability of ChatGPT-generated patient education materials across medical specialties. **BMC Med Inform Decis Mak.** 2024;24:88.
24. Temsah MH, Altamimi I, Algahtani F, et al. The role of ChatGPT in patient education and healthcare communication: opportunities and challenges. **Healthcare (Basel).** 2024;12(2):256.
25. Meskó B, Topol EJ. The imperative for artificial intelligence in patient education and health communication. **NPJ Digit Med.** 2023;6:186.

*Corresponding Author: **Zeynel Umut Canbaba**