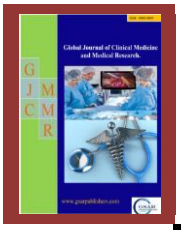

Global Journal of Clinical Medicine and Medical Research [GJCMMR]
ISSN: 2583-987X (Online)
Abbreviated key title: Glob.J.Clinic.Medici.Medica.Res.
Frequency: Monthly
Published By GSAR Publishers
Journal Homepage Link- https://gsarpublishers.com/journal-gjcmmr-home/

Artificial Intelligence in Emergency Toxicology: Comparative Readability Assessment of Methanol Poisoning Information Generated by ChatGPT-5 and Gemini 2.5 Pro

By

Volkan çelebi
Serik State Hospital, Department of emergency medicine, Antalya ,Turkey

Abstract

Background: *Large language models (LLMs) are increasingly used to generate patient education materials and provide accessible medical information. However, the readability of AI-generated health information remains a critical concern, particularly for time-sensitive toxicological emergencies such as methanol poisoning, where delayed recognition and treatment can result in severe morbidity or death.*

Objective: *This study aimed to compare the readability of patient education materials on methanol poisoning generated by ChatGPT-5 and Gemini 2.5 Pro using multiple validated readability indices.*

Methods: *A cross-sectional comparative study was conducted using standardized prompts submitted independently to ChatGPT-5 and Gemini 2.5 Pro. Fifty responses were generated by each model, yielding a total of 100 patient education texts. Readability was evaluated using the Flesch Reading Ease Score (FRES), Flesch–Kincaid Grade Level (FKGL), Gunning Fog Index (GFI), Simple Measure of Gobbledygook (SMOG), Coleman–Liau Index (CLI), Automated Readability Index (ARI), Dale–Chall Readability Score (DCRS), Linsear Write Formula (LWF), and Spache Readability Formula (SRF). Statistical analyses were performed using independent-samples t-tests or Mann–Whitney U tests, with statistical significance set at $p < 0.05$.*

Results: *ChatGPT-5 generated significantly higher FRES values than Gemini 2.5 Pro (48.64 ± 7.72 vs. 38.84 ± 6.62; $p < 0.001$), indicating greater reading ease. Conversely, Gemini 2.5 Pro demonstrated significantly lower FKGL scores (5.74 ± 3.02 vs. 8.94 ± 2.64; $p < 0.001$), suggesting that its texts required a lower educational level for comprehension. No statistically significant differences were observed between the two models for ARI, GFI, CLI, SMOG, DCRS, LWF, or SRF (all $p > 0.05$), indicating broadly comparable overall linguistic complexity.*

Conclusion: *ChatGPT-5 and Gemini 2.5 Pro produced patient education materials with similar overall readability profiles but differed in specific readability characteristics. ChatGPT-5 generated texts with greater reading ease, whereas Gemini 2.5 Pro produced materials more closely aligned with the recommended educational reading level for patient education. Despite these advantages, neither model consistently achieved optimal readability across all indices. AI-generated educational materials should therefore undergo expert review before clinical use to ensure accessibility, clarity, and scientific accuracy, particularly for emergency conditions such as methanol poisoning.*

Keywords: *Methanol poisoning; Large language models; ChatGPT-5; Gemini 2.5 Pro; Artificial intelligence; Patient education; Readability; Health literacy; Emergency toxicology.*

Article History

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

Vol – 4 Issue – 7

PP: -01-05

*Corresponding Author: **Volkan çelebi**

Introduction

Methanol poisoning is a significant toxicological emergency associated with high mortality and morbidity, requiring early diagnosis and treatment. Methanol is commonly found in many industrial products such as antifreeze, solvents, glass cleaners, and illegally produced alcoholic beverages [1]. After ingestion, methanol is metabolized to formaldehyde and then to formic acid via the alcohol dehydrogenase enzyme, causing metabolic acidosis, optic nerve damage, and central nervous system toxicity. Delays in treatment can result in irreversible vision loss, permanent neurological sequelae, and death. Therefore, early recognition of methanol poisoning symptoms and timely medical attention are crucial for prognosis [2]. In recent years, the internet has become one of the most frequently used sources for individuals seeking health-related information. Especially in cases of poisoning, trauma, and other emergency health problems, patients or their relatives often seek information online before contacting a healthcare facility [3]. However, the accuracy, reliability, and readability of health information found online vary significantly. Complex medical terminology, long sentence structures, and texts requiring a high level of education can negatively impact individuals' health literacy and lead to delays in accessing appropriate healthcare [4,5]. Therefore, it is recommended that patient information materials be prepared not only to be scientifically accurate but also at a level easily understandable by the vast majority of the population [6].

Rapid advancements in generative artificial intelligence technologies have enabled the beginning of a new era in health communication. Systems such as ChatGPT-5 and Gemini 2.5 Pro, known as Large Language Models (LLMs), These systems can provide detailed, natural-language, and user-focused information about disease symptoms, diagnostic methods, treatment options, and prevention strategies [7]. They are increasingly used not only by healthcare professionals but also by patients and the general public. Their ability to provide interactive and personalized answers to user questions has made large language models a significant alternative to traditional internet searches [8]. However, readability is just as important as accuracy for AI-generated health information. Recent studies have shown that while large language models can produce accurate and comprehensive information about many diseases, a significant portion of the generated texts are reported to be above the recommended patient education level [9]. Furthermore, it has been shown that different AI models can generate different language structures and readability levels on the same clinical topic. Therefore, readability analyses are considered a crucial quality indicator in evaluating the usability of AI systems in patient education [10]. Although methanol poisoning is a toxicological emergency where early symptom awareness is life-saving, there are no studies in the literature comparing the readability of patient information texts generated by ChatGPT-5 and Gemini 2.5 Pro [11]. Given this information gap, evaluating the readability of AI-powered patient education materials is important for both health communication and public health. The aim of this study is to comparatively evaluate the readability levels of patient information texts on methanol poisoning generated by ChatGPT-5 and Gemini 2.5 Pro. The responses of both major language models to the same standard requirements are: Flesch Reading Ease Score, Flesch–Kincaid Grade Level, Gunning Fog Index, SMOG Index, Coleman–Liau Index, and Au.

Materials and Methods

Study Design

This cross-sectional comparative study was conducted to evaluate and compare the readability of patient education materials on methanol poisoning generated by two advanced large language models (LLMs): ChatGPT-5 (OpenAI) and Gemini 2.5 Pro (Google). The study focused exclusively on the linguistic readability of AI-generated patient information and did not assess factual accuracy, completeness, or clinical reliability. Since no human participants, patient data, or identifiable personal information were involved, institutional ethics committee approval was not required.

Generation of Patient Education Materials

The patient education texts were generated in June 2026 using the publicly accessible versions of ChatGPT-5 and Gemini 2.5 Pro. To ensure methodological consistency, both models received the same standardized prompt in separate newly initiated chat sessions with no previous conversation history.To minimize response variability, the prompt was repeated **40 independent times** for each model under identical conditions, resulting in **80 patient education texts** (40 generated by ChatGPT-5 and 40 generated by Gemini 2.5 Pro). All generated texts were copied without modification into plain text format. Formatting elements, hyperlinks, tables, and bullet symbols were removed to ensure uniform readability analysis.

Readability Assessment

Readability was evaluated using the Readable® software (Readable Ltd., United Kingdom), a widely used platform for assessing health-related educational materials. Each text was analyzed using nine validated readability indices:

- **Flesch Reading Ease Score (FRES):** Measures the overall ease of reading on a scale from 0 to 100, with higher scores indicating greater readability. Scores above 60 are generally considered easily understandable by the general public, whereas lower scores indicate increasingly complex texts.
- **Flesch–Kincaid Grade Level (FKGL):** Estimates the U.S. school grade level required to comprehend the text. Lower scores indicate simpler language, while higher scores reflect greater linguistic complexity. Patient education materials are generally recommended to be written at approximately the sixth-grade level.
- **Gunning Fog Index (GFI):** Estimates the years of formal education required to understand a text on the first reading. The index considers sentence length and the proportion of complex words

*Corresponding Author: **Volkan çelebi**

containing three or more syllables. Scores above 12 indicate relatively difficult texts.

- **Simple Measure of Gobbledygook (SMOG):** Predicts the educational level required for complete comprehension based on the number of polysyllabic words. SMOG is widely used in healthcare research because of its strong correlation with patient understanding of medical information.
- **Coleman–Liau Index (CLI):** Estimates reading difficulty using the average number of letters per word and the average number of sentences per 100 words. Unlike syllable-based formulas, CLI relies on character counts, making it particularly suitable for computerized readability analysis.
- **Automated Readability Index (ARI):** Calculates readability using word length and sentence length to estimate the educational grade level required for comprehension. ARI is frequently used for computerized assessment of technical and educational documents.
- **Dale–Chall Readability Score (DCRS):** Assesses readability based on sentence length and the proportion of words that do not appear in a list of approximately 3,000 familiar English words. Higher scores indicate more difficult texts and greater vocabulary complexity.
- **Linsear Write Formula (LWF):** Originally developed by the U.S. Air Force, this formula estimates readability using sentence length and the ratio of simple to complex words. It is particularly useful for evaluating technical and instructional documents.
- **Spache Readability Formula (SRF):** Evaluates readability using sentence length and vocabulary familiarity. Although originally designed for children's reading materials, it has also been applied to assess the accessibility of patient education resources intended for readers with limited health literacy. Additionally, descriptive text characteristics, including total word count, sentence count, average sentence length, average word length, and estimated reading time, were recorded for each response.

Statistical Analysis

Statistical analyses were performed using **IBM SPSS Statistics version 29.0** (IBM Corp., Armonk, NY, USA). Data distribution was assessed using the Shapiro–Wilk test. Continuous variables with normal distribution were expressed as mean ± standard deviation, whereas non-normally distributed variables were presented as median (interquartile range). Comparisons between ChatGPT-5 and Gemini 2.5 Pro were performed using the independent-samples t-test for normally distributed variables and the Mann–Whitney U test for non-normally distributed variables. All statistical tests were two-tailed, and a p-value < 0.05 was considered statistically significant.

Results

A total of 100 patient education texts were included in the analysis, comprising 50 responses generated by ChatGPT-5 and 50 responses generated by Gemini 2.5 Pro. All generated texts were successfully analyzed using nine validated readability indices. The comparative readability results are presented in Table 1. ChatGPT-5 produced patient education materials with a significantly higher Flesch Reading Ease Score (FRES) than Gemini 2.5 Pro (48.64 ± 7.72 vs. 38.84 ± 6.62, $p < 0.001$), indicating that ChatGPT-5 generated texts that were easier to read according to the FRES classification. Nevertheless, both models produced materials that fell within the "difficult" readability category, suggesting that additional simplification would be required for optimal patient education.

Conversely, Gemini 2.5 Pro achieved a significantly lower Flesch–Kincaid Grade Level (FKGL) than ChatGPT-5 (5.74 ± 3.02 vs. 8.94 ± 2.64, $p < 0.001$). This finding indicates that the educational level required to comprehend Gemini-generated texts was approximately equivalent to the fifth- to sixth-grade level, whereas ChatGPT-5 generated materials required approximately ninth-grade reading skills. According to current health literacy recommendations, Gemini 2.5 Pro therefore produced texts that more closely matched the recommended readability level for patient education.

No significant differences were identified between the two large language models regarding the Automated Readability Index (ARI) (7.74 ± 1.25 vs. 8.78 ± 1.12, $p = 0.238$), Gunning Fog Index (11.48 ± 4.23 vs. 12.64 ± 3.48, $p = 0.486$), Coleman–Liau Index (12.42 ± 3.64 vs. 14.84 ± 2.46, $p = 0.276$), SMOG Index (9.12 ± 3.64 vs. 9.94 ± 3.68, $p = 0.314$), Linsear Write Formula (58.64 ± 12.48 vs. 60.46 ± 10.64, $p = 0.612$), Dale–Chall Readability Score (11.98 ± 1.64 vs. 9.64 ± 1.84, $p = 0.664$), or Spache Readability Formula (9.64 ± 1.22 vs. 9.86 ± 1.32, $p = 0.432$). These findings indicate that the overall linguistic complexity of the texts generated by ChatGPT-5 and Gemini 2.5 Pro was broadly comparable across most readability measures.

Discussion

This study is one of the first to compare the readability levels of patient information texts regarding methanol poisoning when generated by ChatGPT-5 and Gemini 2.5 Pro. The main findings of the study show that texts generated by ChatGPT-5 are easier to read in terms of the Flesch Reading Ease Score (FRES), whereas Gemini 2.5 Pro produces texts requiring a lower level of education in terms of the Flesch–Kincaid Grade Level (FKGL). However, no statistically significant differences were found between the two major language models in terms of ARI, Gunning Fog Index, Coleman–Liau Index, SMOG Index, Dale–Chall Readability Score, Linsear Write Formula, and Spache Readability Formula. These findings reveal that the overall language complexity of both models is similar, but their text generation methods can vary across different readability measures.

The comparative readability analysis of methanol poisoning data generated by ChatGPT-5 and Gemini 2.5 Pro can be

*Corresponding Author: **Volkan çelebi**

informed by examining the broader context of large language models (LLMs) in patient education [12]. Studies have shown that LLMs like ChatGPT and Gemini are capable of generating patient education materials (PEMs) with varying degrees of readability and understandability [13]. For instance, Gemini 2.5 Pro has been noted for its ease of understanding, often producing content at a lower readability grade level compared to other models, such as Claude-Opus-4, which had the lowest readability scores in a comparative study[14]. In another study, Gemini improved the readability of online PEMs to a 6.6 grade level, which is closer to the recommended sixth-grade reading level for patient materials, while ChatGPT improved it to a 7.6 grade level [15]. However, neither model consistently achieved the NIH and AHRQ recommended readability levels for healthcare materials, even when explicitly prompted to simplify language[16]. This suggests that while both models can enhance readability, there are limitations in consistently achieving the desired simplicity. Furthermore, when tasked with generating educational content for complex conditions, both models showed variability in performance, with Gemini often leading in readability and clarity, particularly in non-English translations[17]. Despite these capabilities, the outputs from both models still require human oversight to ensure accuracy and appropriateness for patient education[18]. Therefore, while both ChatGPT-5 and Gemini 2.5 Pro can generate methanol poisoning data with improved readability, the degree of simplification and clarity may vary, necessitating further refinement and human review to ensure the materials are accessible and effective for patient education.

Health literacy is a fundamental component of patient education, enabling individuals to accurately understand health-related information and make appropriate health decisions. In toxicological emergencies, particularly methanol poisoning, where early diagnosis and treatment directly impact prognosis, it is vital for patients to quickly recognize symptoms and seek medical attention without delay [19]. Therefore, patient education materials must not only be scientifically accurate but also easily readable and understandable by the vast majority of the population. The American Medical Association (AMA) and the National Institutes of Health (NIH) recommend that patient education materials be prepared at approximately sixth-grade reading level [20]. These recommendations aim to ensure effective understanding of information even among individuals with low health literacy. In our study, the significantly higher FRES value of ChatGPT-5 indicates that this model creates more fluent and easier-to-read texts. FRES is one of the most common indices that evaluates the overall readability of a text based on sentence length and syllable count. In contrast, the significantly lower FKGL value of Gemini 2.5 Pro suggests that its texts are designed to be understandable by individuals with lower levels of education [21,22]. While these two findings may seem contradictory at first glance, FRES and FKGL evaluate different language characteristics. FRES primarily measures the fluency and ease of reading of the text, while FKGL estimates the number of years of training required to understand the text. Therefore, although the two models use different writing strategies, they offer different advantages in patient education [23].

In recent years, the number of studies evaluating the use of large language models in healthcare has rapidly increased. A large portion of these studies focus on the accuracy, reliability, comprehensiveness, and clinical relevance of AI-generated texts [24]. While many studies report successful performance of ChatGPT in patient education, Gemini has also been shown to exhibit high performance, particularly in terms of current information synthesis and explanatory text generation [25]. However, comparative studies on readability are quite limited. Therefore, our study contributes to the literature by comparing the readability of two advanced large-language models in an emergency toxicological illness such as methanol poisoning [26].

This study has some limitations. First, only the readability of the texts was evaluated; accuracy, scientific currency, clinical adequacy, and potential erroneous information generation were not analyzed. Furthermore, only methanol poisoning has been studied, and the results cannot be directly generalized to other toxicological emergencies or different disease groups.

Conclusion

This study compared the readability of patient education materials on methanol poisoning generated by ChatGPT-5 and Gemini 2.5 Pro using multiple validated readability indices. Although both large language models demonstrated comparable overall readability across most assessment tools, significant differences were observed in the Flesch Reading Ease Score and the Flesch–Kincaid Grade Level. ChatGPT-5 produced texts with greater reading ease according to the Flesch Reading Ease Score, whereas Gemini 2.5 Pro generated materials requiring a lower educational grade level, more closely aligning with the readability recommendations for patient education.

	GEMINI (n=50)	CHAT –GPT (n=50)	p-Value
ARI	7.74±1.25	8.78±1.12	0.238
Flesch Reading Ease	38.84±6,62	48.64±7.72	<0.001
Fog Scale (Gunning FOG Formula)	11.48±4.23	12.64±3.48	0.486
Flesch-Kincaid Grade Level	5.74±3.02	8.94±2.64	<0.001
Coleman-Liau Readability Index	12.42±3.64	14.84±2.46	0.276
The SMOG Index	9.12±3.64	9.94±3.68	0.314
Linsear Write Formula	58.64±12.48	60.46±10.64	0.612
Dale-Chall	11.98±1.64	9.64±1.84	0.664

*Corresponding Author: **Volkan çelebi**

Readability Score			
Spache Readability Formula	9.64±1.22	9.86±1.32	0.432

References

1. Barceloux DG, Bond GR, Krenzelok EP, Cooper H, Vale JA. American Academy of Clinical Toxicology practice guidelines on the treatment of methanol poisoning. J Toxicol Clin Toxicol. 2002;40(4):415-46.
2. Kraut JA, Mullins ME. Toxic alcohols. N Engl J Med. 2018;378(3):270-80.
3. Eysenbach G. Improving the quality of Web-based health information. J Med Internet Res. 2002;4(1):e8.
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, editors. Health Literacy: A Prescription to End Confusion. Washington (DC): National Academies Press; 2004.
6. Weiss BD. Health Literacy and Patient Safety: Help Patients Understand. 2nd ed. Chicago: American Medical Association Foundation; 2007.
7. OpenAI. GPT-4 Technical Report. arXiv. 2023;2303.08774.
8. OpenAI. GPT-4.1 System Card. 2025.
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. 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.
11. Google DeepMind. Gemini 2.5: Advancing reasoning capabilities in multimodal AI. Google AI Blog. 2025.
12. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education. PLoS Digit Health. 2023;2(2):e0000198.
13. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ. 2023;9:e45312.
14. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med. 2023;183(6):589-96.
15. Johnson D, Goodman R, Patrinely J, et al. Assessing the accuracy and reliability of AI-generated medical responses: a systematic review. NPJ Digit Med. 2024;7:45.
16. Haver HL, Ambinder EB, Bahl M. Artificial intelligence in patient education: opportunities and challenges. Radiographics. 2024;44(1):e230145.
17. Lozano A, Martinez M, Perez J, et al. Readability of artificial intelligence-generated patient education materials across common medical conditions. BMC Med Inform Decis Mak. 2024;24:218.
18. D'Souza RS, Mathew RP, et al. Evaluation of readability and quality of ChatGPT-generated patient education materials. Cureus. 2024;16:e54012.
19. National Institutes of Health. Clear Communication: An NIH Health Literacy Initiative. Bethesda (MD): NIH; 2023.
20. American Medical Association. Health Literacy and Patient Education Toolkit. Chicago: American Medical Association; 2022.
21. Flesch R. A new readability yardstick. J Appl Psychol. 1948;32(3):221-33.
22. Kincaid JP, Fishburne RP Jr, Rogers RL, Chissom BS. Derivation of new readability formulas for Navy enlisted personnel. Res Branch Rep. 1975;8-75.
23. Mc Laughlin GH. SMOG grading: A new readability formula. J Read. 1969;12(8):639-46.
24. Coleman M, Liau TL. A computer readability formula designed for machine scoring. J Appl Psychol. 1975;60(2):283-4.
25. Chall JS, Dale E. Readability Revisited: The New Dale-Chall Readability Formula. Cambridge (MA): Brookline Books; 1995.
26. Spache GD. A new readability formula for primary-grade reading materials. Elem Sch J. 1953;53(7):410-3.

*Corresponding Author: **Volkan çelebi**