Acta Medica Young Doctors

Orginal Article

Ibrahim Halil Sukur[1], Fesih Ok[2]

1. Department of Urology, Adana City Training and Research Hospital, Adana, Turkey
2. Department of Urology, Adana City Training and Research Hospital, Adana, Turkey

Corresponding Authors: Dr. Ibrahim Halil Sukur, MD, Adress: Adana City Training and Research Hospital, Adana, Turkey

e-Mail: ihalilsukur@gmail.com

Accepted: 2025-10-25

Şükür, İ. H., & Ok, F. (2025). Comparative Readability Analysis of Large Language Model Responses to Stress Urinary Incontinence Questions: ChatGPT versus Gemini. *Acta Medica Young Doctors*, *1*(3), 66–72.

https://doi.org/10.5281/zenodo.17764454

Comparative Readability Analysis of Large Language Model Responses to Stress Urinary Incontinence Questions: ChatGPT versus Gemini

ABSTRACT

Background: Large language models (LLMs) are increasingly used by patients seeking health information. The readability of AI-generated medical content is crucial for patient comprehension and informed decision-making.

Objective: To compare the readability levels of responses generated by ChatGPT and Gemini to stress urinary incontinence (SUI) patient questions using multiple validated readability assessment tools.

Methods: Thirteen commonly asked questions about SUI were posed to both ChatGPT and Gemini. Responses were analyzed using nine standardized readability metrics: Average Reading Level Consensus Calculator (ARLC), Automated Readability Index (ARI), Flesch Reading Ease (FRE), Gunning Fog Index (GFI), Flesch-Kincaid Grade Level (FKGL), Coleman-Liau Index (CLI), SMOG Index, Original Linsear Write Readability Formula, and LINSEAR Write Grade Level Formula. Statistical comparisons were performed using Student's t-test.

Results: ChatGPT responses demonstrated higher complexity across most metrics. The Flesch Reading Ease score was significantly lower for ChatGPT (31.1±14.5) compared to Gemini (43.4±15.4, p=0.047), indicating more difficult text. The LINSEAR Write Grade Level Formula showed a substantial difference, with ChatGPT scoring 20.9±10.9 versus Gemini's 7.9±2.1 (p<0.001). Other indices including ARLC, ARI, GFI, FKGL, and CLI consistently showed higher grade levels for ChatGPT, though differences did not reach statistical significance.

Conclusion: Gemini produced more readable health information about SUI compared to ChatGPT. Both models generated content at grade levels exceeding recommended health literacy standards. These findings highlight the need for readability optimization in AI-generated medical content.

Keywords: Artificial intelligence; Large language models; Health literacy; Readability; Stress urinary incontinence; Patient education

Introduction

The proliferation of artificial intelligence in healthcare has transformed how patients access medical information. Large language models (LLMs) such as OpenAI's ChatGPT and Google's Gemini have emerged as popular tools for health-related queries, offering instant, detailed responses to complex medical questions[1]. However, the utility of these platforms depends critically on the accessibility and comprehensibility of the information they provide.

Health literacy—the ability to obtain, process, and understand basic health information—is essential for informed medical decision-making. Unfortunately, approximately 36% of adults in the United States have limited health literacy, which is associated with poorer health outcomes, decreased medication adherence, and higher healthcare costs[2]. The American Medical Association and National Institutes of Health recommend that patient education materials be written at or below a 6th to 8th-grade reading level to ensure broad accessibility [3].

Stress urinary incontinence (SUI) affects millions of women worldwide and significantly impacts quality of life[4]. Patients with SUI often seek information about their condition, treatment options, and prognosis through various sources, including increasingly through AI chatbots. Understanding the readability of AI-generated responses to SUI-related questions is therefore clinically relevant.

Despite growing interest in LLM applications for health information dissemination, limited research has compared the readability of different AI platforms' outputs in specific medical contexts. This study aimed to assess and compare the readability levels of ChatGPT and Gemini responses to common SUI patient questions using multiple validated readability assessment tools.

Methods

Question Development

Thirteen frequently asked questions about stress urinary incontinence were compiled based on clinical experience and patient education guidelines. The questions covered etiology, diagnosis, non-surgical treatments, surgical options, risks, outcomes, and recovery expectations.

AI Response Generation

Identical question prompts were submitted to ChatGPT (OpenAI) and Gemini (Google) in December 2024. Responses were collected verbatim without any modifications to preserve authentic AI output.

Readability Assessment

Nine validated readability metrics were employed to comprehensively assess text complexity[5]:

Average Reading Level Consensus Calculator (ARLC) - aggregates multiple formulas

Automated Readability Index (ARI) - based on characters per word and words per sentence

Flesch Reading Ease (FRE) - inversely related to difficulty (0-100 scale)

Gunning Fog Index (GFI) - estimates years of formal education needed

Flesch-Kincaid Grade Level (FKGL) - U.S. school grade level

Coleman-Liau Index (CLI) - based on characters rather than syllables

SMOG Index - estimates years of education needed for comprehension

Original Linsear Write Readability Formula - originally designed for technical writing

LINSEAR Write Grade Level Formula - grade-level adaptation of Linsear

Statistical Analysis

Descriptive statistics were calculated for all readability scores from both language models across the nine scales. The Shapiro-Wilk test was performed to assess the normality of data distribution. As the data followed a normal distribution, results are presented as mean ± standard deviation. Comparisons between the two language models (ChatGPT and Gemini) were conducted using Student's t-test for independent samples. A p-value of less than 0.05 was considered statistically significant. All statistical analyses were performed using IBM SPSS Statistics version 27 (IBM Corp., Armonk, USA).

Results

Readability assessments of responses to stress urinary incontinence questions from two language models (ChatGPT and Gemini) were conducted using nine different readability scales. The mean and standard deviation for each index, as well as corresponding p-values, are presented in Table 1.

ChatGPT responses generally had higher grade level scores, indicating more complex language, than Gemini. The Average Reading Level Consensus Calc (ARLC) was 12.4±2.1 for ChatGPT and 11.5±1.9 for Gemini (p=0.255). The Automated Readability Index (ARI) was 12.6±3.5 for ChatGPT and 11.5±2.4 for Gemini (p=0.344). The Flesch Reading Ease (FRE) score, which inversely correlates with text difficulty (higher scores mean easier text), was lower in ChatGPT (31.1±14.5) compared to Gemini (43.4±15.4), with this difference reaching statistical significance (p=0.047).

For the Gunning Fog Index (GFI), ChatGPT scored 13.8±2.7 compared to Gemini's 12.9±2.9 (p=0.462). The Flesch-Kincaid Grade Level (FKGL) was 11.7±2.6 for ChatGPT and 10.3±2.3 for Gemini (p=0.173). ChatGPT had a higher Coleman-Liau Index (CLI) at 15.7±2.8 compared to Gemini's 13.6±2.7, though this difference approached but did not reach statistical significance (p=0.067). The SMOG Index was similar between ChatGPT (9.6±1.8) and Gemini (9.4±1.7), with p=0.794.

The Original Linsear Write Readability Formula showed similar scores for both ChatGPT (73.9±13.9) and Gemini (73.4±10.5), with no significant difference (p=0.912). However, the LINSEAR Wrıte Grade Level Formula yielded a substantially higher score for ChatGPT (20.9±10.9) compared to Gemini (7.9±2.1), and this difference was statistically significant (p<0.001)..

Discussion

This study provides the first comparative analysis of readability between ChatGPT and Gemini for patient-oriented medical information about stress urinary incontinence. Our findings reveal that while both platforms generate comprehensive responses, Gemini consistently produced more accessible content than ChatGPT, with statistically significant differences observed in two key metrics.

The Flesch Reading Ease score difference (31.1 vs. 43.4, p=0.047) is clinically meaningful. ChatGPT's FRE score indicates "difficult" college-level text, while Gemini's score suggests "fairly difficult" high school-level material. Neither platform achieved the recommended 60-70 FRE score associated with 8th to 9th-grade readability—the target for general patient education materials [3, 6]. However, Gemini's superior performance suggests its language generation algorithms may be better calibrated for patient-facing communication.

The dramatic difference in LINSEAR Write Grade Level scores (20.9 vs. 7.9, $p<0.001$) warrants particular attention. This formula specifically evaluates sentence structure and word complexity in instructional text. ChatGPT's extremely high score (above college level) suggests the use of complex sentence constructions and technical vocabulary that may challenge patients with limited health literacy. Gemini's score, approximating 8th-grade level, aligns more closely with health literacy recommendations[3].

Several factors may contribute to these readability differences. First, ChatGPT's responses employed more formal medical terminology and complex syntactic structures. For example, phrases like "autologous fascial sling" and "urodynamics" appeared without simplified explanations. Second, ChatGPT utilized longer sentences with multiple dependent clauses, increasing cognitive load[7]. Third, ChatGPT provided extensive disclaimers about its limitations, which, while ethically important, added complexity.

Conversely, Gemini incorporated visual elements (emoji indicators), shorter paragraphs, bulleted lists, and more conversational language. These formatting choices, beyond mere linguistic simplicity, likely contributed to improved readability scores[8]. Gemini also employed analogies and explanations (e.g., describing the sling as "acting like a hammock"), which can enhance comprehension for non-expert readers.

The consistency of ChatGPT's higher grade-level scores across multiple independent metrics (ARLC, ARI, GFI, FKGL, CLI) strengthens our conclusions, despite some not reaching statistical significance. This pattern suggests a systematic difference in content generation approaches rather than random variation.

These findings have important implications for patient care. Patients with SUI are predominantly middle-aged and older women, populations with documented health literacy challenges[2, 9]. When AI-generated information exceeds patients' reading comprehension, it may lead to misunderstanding of treatment options, unrealistic expectations, non-adherence to recommendations, or avoidance of seeking professional care. Healthcare providers who recommend or allow patients to use AI tools should be aware of these readability disparities[10].

Previous studies have documented similar concerns. Research by Ayers et al. [11] found that while ChatGPT provides empathetic responses to patient questions, the quality and appropriateness varied. Johnson et al. [8] evaluated ChatGPT's responses to common oncology questions and found readability levels consistently exceeded the 8th-grade benchmark. Our study extends this literature by including comparative analysis with Gemini, a newer competitor in the LLM space, and by focusing on a specific urological condition.

Several strategies could improve AI-generated health content readability. Developers could implement post-generation readability filtering that prompts simplification when grade levels exceed thresholds. Training datasets could prioritize patient-facing educational materials over technical medical literature [12]. Users could be prompted to specify their preferred reading level. Healthcare institutions implementing AI chatbots could fine-tune models using simplified medical terminology databases[13].

Limitations

This study has several limitations. First, we examined only one medical condition (SUI);

readability patterns may differ across other health topics. Second, we used a specific set of 13 questions; alternative phrasings might yield different results. Third, readability formulas have limitations—they cannot assess conceptual difficulty, accuracy, or cultural appropriateness[14, 15]. Fourth, we did not evaluate actual patient comprehension or satisfaction. Fifth, AI models are continuously updated; our findings reflect specific versions tested in December 2024[16]. Finally, we did not account for visual formatting elements that may enhance comprehension beyond measured readability scores.

Conclusion

Gemini generated more readable stress urinary incontinence information compared to ChatGPT, with statistically significant differences in Flesch Reading Ease and LINSEAR Write Grade Level scores. However, both platforms produced content exceeding recommended health literacy standards[3, 6]. As patients increasingly turn to AI for health information, developers must prioritize readability optimization[12]. Healthcare providers should counsel patients about the complexity of AI-generated content and offer supplementary explanations[17]. Future research should evaluate actual patient comprehension, assess additional medical topics, and examine interventions to improve AI health content accessibility.

Acknowledgments

The authors would like to acknowledge the developers of ChatGPT (OpenAI) and Gemini (Google) for providing publicly accessible artificial intelligence platforms that made this comparative study possible.

Declaration of Interest Statement

The authors declare no conflicts of interest related to this study. No financial or personal relationships with organizations or individuals that could inappropriately influence this work exist.

Ethical Approval

Ethical approval was not required for this study as it involved only publicly available artificial intelligence language model responses and did not involve human participants, patient data, or animal subjects.

Data Availability Statement

The complete dataset, including all AI-generated responses and readability scores, is available from the corresponding author upon reasonable request.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. This work was conducted as an independent academic investigation.

uthor Contributions Statement

All authors contributed to the conception and design of the study, analysis and interpretation of data, drafting of the manuscript, critical revision for intellectual content, and final approval of the version to be published. All authors agree to be accountable for all aspects of the work.

Table 1. Comparison of Readability Metrics Between ChatGPT and Gemini Responses to Stress Urinary Incontinence Questions

	ChatGPT	Gemini	P value
Average Reading Level Consensus Calc (ARLC), mean±SD	12.4±2.1	11.5±1.9	0.255
Automated Readability Index (ARI), mean±SD	12.6±3.5	11.5±2.4	0.344
Flesch Reading Ease (FRE), mean±SD	31.1±14.5	43.4±15.4	**0.047**
Gunning Fog Index (GFI), mean±SD	13.8±2.7	12.9±2.9	0.462
Flesch-Kincaid Grade Level (FKGL), mean±SD	11.7±2.6	10.3±2.3	0.173
Coleman-Liau Index (CLI), mean±SD	15.7±2.8	13.6±2.7	0.067
SMOG Index, mean±SD	9.6±1.8	9.4±1.7	0.794
Original Linsear Write Readability Formula, mean±SD	73.9±13.9	73.4±10.5	0.912
LINSEAR Wrıte Grade Level Formula, mean±SD	20.9±10.9	7.9±2.1	**<0.001**

REFERENCES

1. 1. Sezgin, E., J. Sirrianni, and S.L. Linwood, Operationalizing and implementing pretrained, large artificial intelligence linguistic models in the US health care system: outlook of generative pretrained transformer 3 (GPT-3) as a service model. JMIR medical informatics, 2022. **10**(2): p. e32875.

2. 2. Berkman, N.D., et al., Low health literacy and health outcomes: an updated systematic review. Annals of internal medicine, 2011. **155**(2): p. 97-107.

3. 3. Weiss, B.D., Help patients understand. Manual for Clinicians. AMA Foundation, 2007.

4. 4. Lukacz, E.S., et al., Urinary incontinence in women: a review. Jama, 2017. **318**(16): p. 1592-1604.

5. 5. Wang, L.-W., et al., Assessing readability formula differences with written health information materials: application, results, and recommendations. Research in Social and Administrative Pharmacy, 2013. **9**(5): p. 503-516.

6. 6. Paasche-Orlow, M.K., H.A. Taylor, and F.L. Brancati, Readability standards for informed-consent forms as compared with actual readability. New England journal of medicine, 2003. **348**(8): p. 721-726.

7. 7. Sweller, J., Cognitive load theory, learning difficulty, and instructional design. Learning and instruction, 1994. **4**(4): p. 295-312.

8. 8. Shoemaker, S.J., M.S. Wolf, and C. Brach, Development of the Patient Education Materials Assessment Tool (PEMAT): a new measure of understandability and actionability for print and audiovisual patient information. Patient education and counseling, 2014. **96**(3): p. 395-403.

9. 9. Kobayashi, L.C., et al., Aging and functional health literacy: a systematic review and meta-analysis. Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 2016. **71**(3): p. 445-457.

10. 10. Nouri, S.S., et al., Assessing mobile phone digital literacy and engagement in user-centered design in a diverse, safety-net population: mixed methods study. JMIR mHealth and uHealth, 2019. **7**(8): p. e14250.

11. 11. Ayers, J.W., et al., Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA internal medicine, 2023. **183**(6): p. 589-596.

12. 12. Lee, P., S. Bubeck, and J. Petro, Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. New England Journal of Medicine, 2023. **388**(13): p. 1233-1239.

13. 13. Zeng-Treitler, Q., et al., Estimating consumer familiarity with health terminology: a context-based

approach. Journal of the American Medical Informatics Association, 2008. **15**(3): p. 349-356.

14. 14. Haynes, S.N., D. Richard, and E.S. Kubany, Content validity in psychological assessment: A functional approach to concepts and methods. Psychological assessment, 1995. **7**(3): p. 238.

15. 15. Dumenci, L., et al., On the validity of the shortened rapid estimate of adult literacy in medicine (REALM) scale as a measure of health literacy. Communication methods and measures, 2013. **7**(2): p. 134-143.

16. 16. Chen, S., et al., Use of artificial intelligence chatbots for cancer treatment information. JAMA oncology, 2023. **9**(10): p. 1459-1462.

17. 17. DeWalt DA, C.L., Hawk VH, Brach C, Hink A, Health Literacy Universal Precautions Toolkit. 2010.