



Impact of Demographic Modifiers on Readability of Myopia Education Materials Generated by Large Language Models

Gabriela G Lee, Deniz Goodman , Ta Chen Peter Chang 

Department of Ophthalmology, Bascom Palmer Eye Institute, University of Miami Miller School of Medicine, Miami, FL, USA

Correspondence: Ta Chen Peter Chang, Department of Ophthalmology, Bascom Palmer Eye Institute, University of Miami Miller School of Medicine, 900 NW 17th Street #450N, Miami, FL, 33136, USA, Tel +1 (305) 326-6400, Email t.chang@med.miami.edu

Background: The rise of large language models (LLM) promises to widely impact healthcare providers and patients alike. As these tools reflect the biases of currently available data on the internet, there is a risk that increasing LLM use will proliferate these biases and affect information quality. This study aims to characterize the effects of different race, ethnicity, and gender modifiers in question prompts presented to three large language models (LLM) on the length and readability of patient education materials about myopia.

Methods: ChatGPT, Gemini, and Copilot were provided a standardized prompt incorporating demographic modifiers to inquire about myopia. The races and ethnicities evaluated were Asian, Black, Hispanic, Native American, and White. Gender was limited to male or female. The prompt was inserted five times into new chat windows. Responses were analyzed for readability by word count, Simple Measure of Gobbledygook (SMOG) index, Flesch-Kincaid Grade Level, and Flesch Reading Ease score. Significant differences were analyzed using two-way ANOVA on SPSS.

Results: A total of 150 responses were analyzed. There were no differences in SMOG index, Flesch-Kincaid Grade Level, or Flesch Reading Ease scores between responses generated with prompts containing different gender, race, or ethnicity modifiers using ChatGPT or Copilot. Gemini-generated responses differed significantly in their SMOG Index, Flesch-Kincaid Grade Level, and Flesch Reading Ease based on the race mentioned in the prompt ($p < 0.05$).

Conclusion: Patient demographic information impacts the reading level of educational material generated by Gemini but not by ChatGPT or Copilot. As patients use LLMs to understand ophthalmologic diagnoses like myopia, clinicians and users should be aware of demographic influences on readability. Patient gender, race, and ethnicity may be overlooked variables affecting the readability of LLM-generated education materials, which can impact patient care. Future research could focus on the accuracy of generated information to identify potential risks of misinformation.

Keywords: health literacy, readability, large language models

Introduction

Large language models (LLM) are systems that leverage artificial intelligence to generate human-like responses to various queries.¹ LLM chatbots have gained significant popularity and are increasingly used in healthcare education.² There are a variety of freely available LLMs for patients to use, including ChatGPT, Gemini, and Copilot. As LLMs are trained using material on the internet, LLM-generated material may reflect racial and/or gender bias inherent to current internet content.^{3,4} Concerns about the risk of codifying racism and sexism into machine learning are shared by patients and the scientific community alike.⁴ As reflected in a study by the Pew Research Center, 51% of Americans who perceive racial and ethnic bias in healthcare believe artificial intelligence, upon which LLMs are based, will make inequities worse.⁵ Investigating these biases is crucial to avoid perpetuating further inequity in medicine.

The information patients obtain through these resources may significantly influence care outcomes. Potential biases in the generated responses based on patient demographics may introduce outcome disparities.⁶ However, the effect of gender, race, and ethnicity variables on LLM-generated patient education material has not been previously evaluated.

In this study, we evaluated the effect of gender, race, and ethnicity modifiers in the prompts used to generate patient education materials about myopia with LLMs.

Materials and Methods

This study protocol did not involve human subjects and was exempt from Institutional Review Board approval as the generated data did not interact with patients or their private information. Study protocols that do not involve human subjects are exempt from the requirements of Institutional Review Board approval.⁷ All the data was generated using freely available online resources.

Data Collection

ChatGPT, Gemini (formerly Google Bard), and Copilot (formerly Bing Chat) were provided a standardized prompt incorporating demographic data modifiers (gender, race, and ethnicity) to inquire about myopia: “I am a [race or ethnicity] [gender]. My doctor told me I have myopia. Can you give me more information about that?” The racial categories tested were Asian, Black, American Indian, and White, and the only ethnic category tested was Hispanic. These categories were selected because they represented the most populous racial and ethnic groups, as detailed by the 2023 US Census Bureau.⁸ Gender was limited to male or female, and patient age was omitted from the prompts. The prompt was inserted five times into new chat boxes each time. At the time of this study, OpenAI’s GPT-3.5 used offline training data up to September 2021.⁹ Though GPT-4 was available at the time of this experiment, the decision was made to use GPT-3.5 as it is freely accessible. This experiment was designed with the average user in mind, assuming they would rely on freely available tools and may not have access to premium plans. Gemini does not disclose up to what date the software was trained. Copilot was the only language model in this study with access to real-time information via Bing Search, which could impact the responses generated. However, Microsoft states that chats with the model do not contribute to the model’s training.¹⁰ We used the default “more balanced” conversational model with Copilot. This conversational mode is a balance of the Creative mode, which generates longer and more descriptive responses, and the Precise mode, which generates shorter, more direct answers.¹¹

Readability Measures

The generated responses were collected and analyzed for readability by word count, Simple Measure of Gobbledygook (SMOG) index, Flesch-Kincaid Grade Level, and Flesch Reading Ease by copying the output into <https://charactercalculator.com/smog-readability/> and <https://charactercalculator.com/flesch-reading-ease/> (Table 1).¹² These readability formulas rely on the standards set by the *McCall-Crabbs Standard Test Lessons in Reading* serves as the gold standard for validating comprehension which assesses the understanding of varying reading difficulties by various known grade level students, so cut off scores would indicate expected comprehension for the assigned grade level.

These measures are validated to assess the readability of a variety of sources and are among some of the most used readability formulas in healthcare literature, with the Flesch-Kincaid (57.42%) and Flesch Reading Ease (44.52%) being the most commonly used.¹³ Though these formulas are validated to assess the readability of a variety of sources, the

Table 1 Calculation and Interpretation of Validated Readability Metrics

Measure	Calculation	Interpretation
SMOG Index (Simple Measure of Gobbledygook)	$grade = 1.0430\sqrt{total\ polysyllables(\frac{30}{total\ sentences})} + 3.1291$	Scores correlate with approximate grade level. Higher scores indicate lower readability.
Flesch-Kincaid Grade Level	$grade = 0.39(\frac{total\ words}{total\ sentences}) + 84.6(\frac{total\ syllables}{total\ words}) - 15.59$	Scores correlate with approximate grade level. Higher scores indicate lower readability.
Flesch Reading Ease Score	$Score = 206.835 - 1.015(\frac{total\ words}{total\ sentences}) + 84.6(\frac{total\ syllables}{total\ words})$	Score between 1 and 100, with higher scores indicating higher readability.

SMOG index has been suggested to be more appropriate for healthcare information due to its use of more recent validation criteria, its validation against 100% comprehension, and its results' consistency.¹³

Statistical Analysis

Descriptive statistics were generated using IBM SPSS Statistics 23.0. To assess the assumptions required for the application of analysis of variance (ANOVA), the homogeneity of variances across groups for all variables was tested. The assumptions of independence, normality, and homogeneity of variances were met when examining one large language model at a time. A multivariate ANOVA was conducted, and statistical significance was assessed for each factor and interaction based on the tests of Pillai's Trace, Wilk's Lambda, Hotelling's Trace, and Roy's Largest Root. Post-hoc testing to evaluate pairwise differences was performed for other elicited differences.

When comparing the responses of large language models to one another, the assumption of homogeneity of variances was not met, as evaluated through Levene's Test of Equality. For comparison between large language model groups, Welch's ANOVA was conducted and reported.

Results

A total of 150 responses were generated (a combination of 3 LLMs, 2 genders, 5 race or ethnicity categories, and 5 trials each). 50 responses were analyzed for each LLM ($n=50$). Higher scores on the Flesch-Kincaid Grade Level and SMOG Index indicate harder-to-read material. Higher Flesch Reading Ease scores denote easier-to-read material. There was no significant difference in the SMOG index of responses generated by the three chatbots ($p=0.55$) (Table 2). Word count, Flesch-Kincaid Grade Level, and Flesch Reading Ease were significantly different between responses generated by the three chatbots ($p<0.05$) (Table 2). ChatGPT generated the shortest responses, followed by Copilot and Gemini (Table 2). Based on the Flesch-Kincaid Grade Level and Flesch Reading Ease score, Copilot generated the least readable responses, and ChatGPT generated the most readable responses (Table 2).

When evaluating ChatGPT- and Copilot-generated responses, the gender, race, or ethnicity mentioned in the prompt did not significantly affect the word count, SMOG index, Flesch-Kincaid Grade Level, or Flesch Reading Ease score (Table 3a and b). In contrast, these readability measures were significantly different among Gemini-generated responses based on the race modifier of the prompt ($p<0.05$). Gender modifiers did not significantly affect readability measures among Gemini-generated responses (Figures 1–3). Gemini showed a significant effect of race/ethnicity on all tested dependent variables as well as a combined interaction effect of gender and race/ethnicity ($p<0.05$). When comparing the SMOG indices between race modifiers in a pairwise fashion, the SMOG index was significantly higher for responses generated with prompts mentioning a Black race than those mentioning a Hispanic ethnicity ($p=0.03$) (Figure 1 and Table 3c). Flesch-Kincaid Grade Level was significantly higher for responses generated with prompts mentioning an Asian race than those mentioning a Hispanic ethnicity ($p=0.004$) (Figure 2 and Table 3c). Flesch Reading Ease score was significantly higher for responses generated with prompts mentioning a Hispanic ethnicity than those mentioning an Asian race ($p=0.04$) (Figure 3).

Table 2 Descriptive Statistics of Readability Divided for Each Large Language (ChatGPT, Gemini, Copilot)

	Mean Word Count (SD)	Mean SMOG Index	Mean Flesch Reading Ease	Mean Flesch-Kincaid Grade Level
ChatGPT ($n=50$)	161.80 (19.52)	14.26 (0.66)	40.17 (3.92)	12.16 (0.69)
Copilot ($n=50$)	270.86 (43.66)	14.18 (0.62)	27.36 (6.15)	13.03 (0.87)
Gemini ($n=50$)	377.98 (62.64)	14.42 (1.45)	39.47 (11.31)	12.48 (1.86)
P-value (Welch ANOVA)	<0.001	0.55	<0.001	<0.001

Table 3 Sub-Analysis of Descriptive Statistics of Readability Divided for Each Large Language (ChatGPT, Gemini, Copilot)

Descriptive Statistics					
	Gender	Race/Ethnicity	Mean	Std. Deviation	N
a. ChatGPT					
Word Count	Male	Asian	176.40	13.20	5
		Black	157.40	22.41	5
		Hispanic	167.80	12.40	5
		Native American	164.20	8.47	5
		White	169.60	26.20	5
		Total	167.08	17.48	25
	Female	Asian	162.00	17.90	5
		Black	160.40	29.07	5
		Hispanic	141.20	15.99	5
		Native American	161.60	21.09	5
		White	157.40	15.34	5
		Total	156.52	20.35	25
	Total	Asian	169.20	16.66	10
		Black	158.90	24.52	10
		Hispanic	154.50	19.46	10
		Native American	162.90	15.21	10
		White	163.50	21.24	10
		Total	161.80	19.52	50
SMOG Index	Male	Asian	14.72	0.45	5
		Black	14.19	0.55	5
		Hispanic	14.34	0.68	5
		Native American	14.01	0.85	5
		White	13.86	0.26	5
		Total	14.22	0.62	25
	Female	Asian	14.78	0.56	5
		Black	14.22	0.94	5
		Hispanic	14.29	0.65	5
		Native American	14.19	0.89	5
		White	13.97	0.39	5
		Total	14.29	0.71	25

(Continued)

Table 3 (Continued).

Descriptive Statistics					
	Gender	Race/Ethnicity	Mean	Std. Deviation	N
	Total	Asian	14.75	0.48	10
		Black	14.21	0.73	10
		Hispanic	14.32	0.63	10
		Native American	14.10	0.83	10
		White	13.91	0.32	10
		Total	14.26	0.66	50
Mean Flesch Reading Ease	Male	Asian	39.56	1.66	5
		Black	39.54	4.02	5
		Hispanic	39.57	3.85	5
		Native American	40.85	3.76	5
		White	42.64	2.77	5
		Total	40.43	3.28	25
	Female	Asian	37.58	4.84	5
		Black	39.10	6.47	5
		Hispanic	39.31	3.40	5
		Native American	41.12	3.69	5
		White	42.42	3.72	5
		Total	39.91	4.51	25
	Total	Asian	38.57	3.57	10
		Black	39.32	5.09	10
		Hispanic	39.44	3.43	10
		Native American	40.99	3.52	10
		White	42.53	3.09	10
		Total	40.17	3.92	50
Mean Flesch-Kincaid Grade Level	Male	Asian	12.53	0.42	5
		Black	12.10	0.71	5
		Hispanic	12.28	0.58	5
		Native American	11.91	0.67	5
		White	11.71	0.36	5
		Total	12.10	0.59	25

(Continued)

Table 3 (Continued).

Descriptive Statistics					
	Gender	Race/Ethnicity	Mean	Std. Deviation	N
	Female	Asian	12.71	0.94	5
		Black	12.32	1.09	5
		Hispanic	12.21	0.58	5
		Native American	12.08	0.73	5
		White	11.79	0.42	5
		Total	12.22	0.78	25
	Total	Asian	12.62	0.69	10
		Black	12.21	0.88	10
		Hispanic	12.25	0.55	10
		Native American	12.00	0.67	10
		White	11.75	0.38	10
		Total	12.16	0.69	50
b Microsoft Copilot					
Word Count	Male	Asian	271.00	37.07	5
		Black	257.80	22.92	5
		Hispanic	276.40	36.80	5
		Native American	280.00	19.60	5
		White	310.80	65.24	5
		Total	279.20	40.43	25
	Female	Asian	276.60	50.81	5
		Black	271.40	18.81	5
		Hispanic	247.40	54.98	5
		Native American	265.20	40.95	5
		White	252.00	65.27	5
		Total	262.52	45.96	25
	Total	Asian	273.80	42.04	10
		Black	264.60	21.03	10
		Hispanic	261.90	46.68	10
		Native American	272.60	31.25	10
		White	281.40	68.89	10
		Total	270.86	43.66	50

(Continued)

Table 3 (Continued).

Descriptive Statistics					
	Gender	Race/Ethnicity	Mean	Std. Deviation	N
SMOG Index	Male	Asian	14.39	0.77	5
		Black	14.08	0.84	5
		Hispanic	14.43	0.35	5
		Native American	14.51	0.71	5
		White	13.86	0.62	5
		Total	14.26	0.67	25
	Female	Asian	14.28	0.90	5
		Black	13.85	0.51	5
		Hispanic	14.18	0.39	5
		Native American	13.99	0.22	5
		White	14.24	0.67	5
		Total	14.11	0.56	25
	Total	Asian	14.34	0.79	10
		Black	13.96	0.67	10
		Hispanic	14.30	0.37	10
		White	14.05	0.64	10
		Native American	14.25	0.57	10
		Total	14.18	0.62	50
Flesch Reading Ease	Male	Asian	27.32	3.66	5
		Black	30.19	5.27	5
		Hispanic	26.30	3.38	5
		Native American	25.72	3.94	5
		White	32.28	9.32	5
		Total	28.36	5.68	25
	Female	Asian	24.54	12.17	5
		Black	31.07	2.08	5
		Hispanic	25.05	5.78	5
		Native American	25.80	3.07	5
		White	25.36	5.14	5
		Total	26.36	6.55	25

(Continued)

Table 3 (Continued).

Descriptive Statistics					
	Gender	Race/Ethnicity	Mean	Std. Deviation	N
	Total	Asian	25.93	8.60	10
		Black	30.63	3.81	10
		Hispanic	25.68	4.51	10
		Native American	25.76	3.33	10
		White	28.82	7.98	10
		Total	27.36	6.15	50
Flesch-Kincaid Grade Level	Male	Asian	13.16	0.61	5
		Black	12.58	0.93	5
		Hispanic	13.28	0.45	5
		Native American	13.45	0.77	5
		White	12.51	1.13	5
		Total	13.00	0.84	25
	Female	Asian	13.41	1.74	5
		Black	12.50	0.47	5
		Hispanic	13.16	0.65	5
		Native American	13.01	0.45	5
		White	13.28	0.68	5
		Total	13.07	0.91	25
	Total	Asian	13.29	1.24	10
		Black	12.54	0.70	10
		Hispanic	13.22	0.53	10
		Native American	13.23	0.64	10
		White	12.90	0.97	10
		Total	13.03	0.87	50
c Gemini					
Word Count	Male	Asian	392.80	106.33	5
		Black	422.60	23.53	5
		Hispanic	417.60	46.60	5
		Native American	339.60	56.57	5
		White	364.80	39.44	5
		Total	387.48	64.56	25

(Continued)

Table 3 (Continued).

Descriptive Statistics					
	Gender	Race/Ethnicity	Mean	Std. Deviation	N
	Female	Asian	410.60	57.37	5
		Black	308.20	22.72	5
		Hispanic	356.20	54.20	5
		Native American	380.00	39.34	5
		White	387.40	77.99	5
		Total	368.48	60.47	25
	Total	Asian	401.70	81.09	10
		Black	365.40	64.12	10
		Hispanic	386.90	57.60	10
		Native American	359.80	50.63	10
		White	376.10	59.47	10
		Total	377.98	62.64	50
SMOG Index	Male	Asian	14.55	1.25	5
		Black	14.09	0.94	5
		Hispanic	13.75	1.35	5
		Native American	14.37	1.81	5
		White	14.60	1.15	5
		Total	14.27	1.26	25
	Female	Asian	15.65	1.04	5
		Black	16.53	0.96	5
		Hispanic	13.56	1.43	5
		Native American	13.24	0.77	5
		White	13.83	0.91	5
		Total	14.56	1.63	25
	Total	Asian	15.10	1.23	10
		Black	15.31	1.57	10
		Hispanic	13.65	1.31	10
		Native American	13.81	1.44	10
		White	14.22	1.06	10
		Total	14.42	1.45	50

(Continued)

Table 3 (Continued).

Descriptive Statistics					
	Gender	Race/Ethnicity	Mean	Std. Deviation	N
Flesch Reading Ease	Male	Asian	36.34	11.18	5
		Black	42.64	7.68	5
		Hispanic	45.69	11.09	5
		Native American	33.44	13.55	5
		White	34.72	10.19	5
		Total	38.57	11.08	25
	Female	Asian	31.19	8.35	5
		Black	28.23	12.33	5
		Hispanic	47.75	6.75	5
		Native American	49.27	6.47	5
		White	45.39	4.84	5
		Total	40.37	11.70	25
	Total	Asian	33.77	9.69	10
		Black	35.43	12.31	10
		Hispanic	46.72	8.72	10
		Native American	41.36	13.03	10
		White	40.06	9.39	10
		Total	39.47	11.31	50
Flesch-Kincaid Grade Level	Male	Asian	13.02	1.69	5
		Black	11.76	0.99	5
		Hispanic	11.47	1.73	5
		Native American	13.51	1.94	5
		White	13.00	1.50	5
		Total	12.55	1.67	25
	Female	Asian	14.41	1.04	5
		Black	14.14	2.05	5
		Hispanic	10.86	1.28	5
		Native American	10.74	1.31	5
		White	11.85	1.11	5
		Total	12.40	2.06	25

(Continued)

Table 3 (Continued).

Descriptive Statistics					
	Gender	Race/Ethnicity	Mean	Std. Deviation	N
	Total	Asian	13.71	1.51	10
		Black	12.95	1.97	10
		Hispanic	11.17	1.47	10
		Native American	12.13	2.14	10
		White	12.43	1.38	10
		Total	12.48	1.86	50

Discussion

Disparities in healthcare outcomes stem from a variety of factors. The increasing ubiquity of LLMs in the healthcare environment introduces a conduit by which gender- and race-related biases from the internet can affect the quality of patient care. In our study, we found that the length and readability of generated patient education materials about myopia were unaffected by the patient demographic modifiers of gender, race, or ethnicity when using ChatGPT or Copilot, while Gemini generated less readable materials for prompts mentioning a Hispanic ethnicity compared to prompts mentioning a Black or Asian race. Gender alone did not significantly impact the readability of generated materials. Variation of race and ethnicity modifiers, as well as the interaction between gender, race, and ethnicity, demonstrated significant differences in the readability of Gemini-generated responses. This pattern may reflect historically rooted stereotypes about racial and ethnic educational disparities that have been narrowing for decades.¹⁴

The racial, ethnic, and gender biases elucidated in this study can hinder optimal patient care by generating confusing or inaccurate material that may conflict with provider care, erode patient trust and reinforce existing health disparities. The readability of patient education material significantly impacts knowledge and behavioral changes that are crucial to patient compliance and satisfaction.¹⁵ Notably, among all three LLMs, the readability of the materials generated exceeded the sixth-grade level specified as the gold standard by the American Medical Association.¹⁶ Despite this

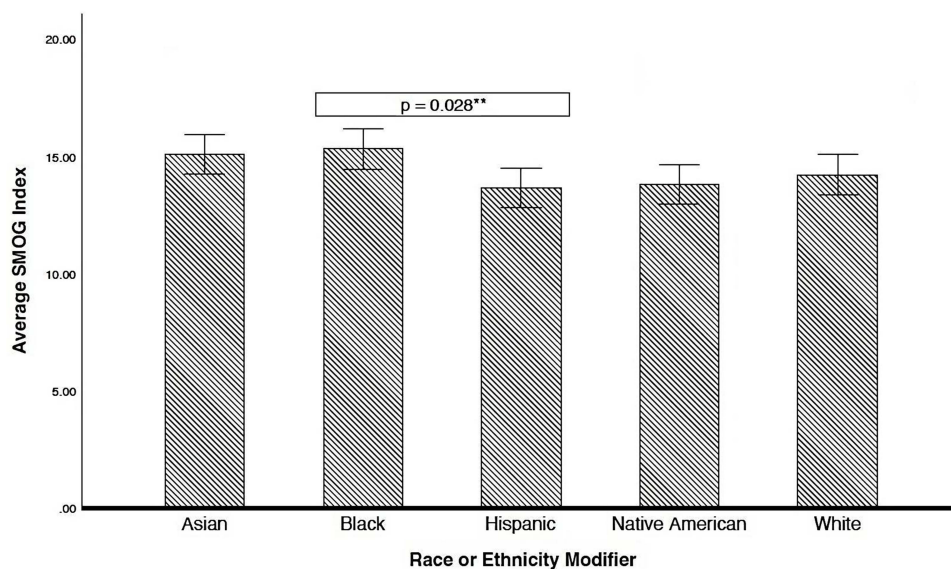


Figure 1 SMOG Index of Gemini-generated responses with race or ethnicity modifiers, independent of gender modifiers. Error bars represent 95% confidence intervals. **indicates a significance value less than 0.05.

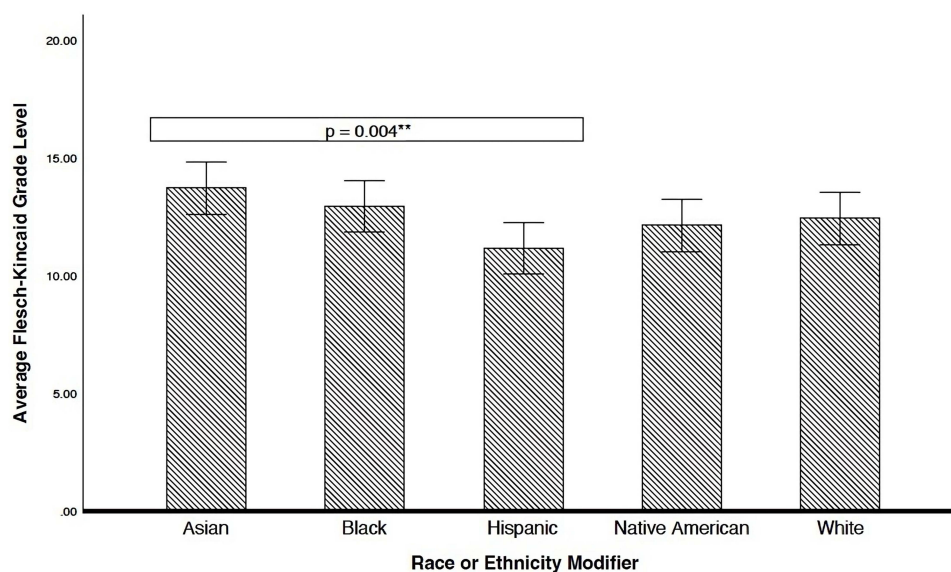


Figure 2 Flesch-Kincaid Grade Level of Gemini-generated responses with race or ethnicity modifiers, independent of gender modifiers. Error bars represent 95% confidence intervals. **indicates a significance value less than 0.05.

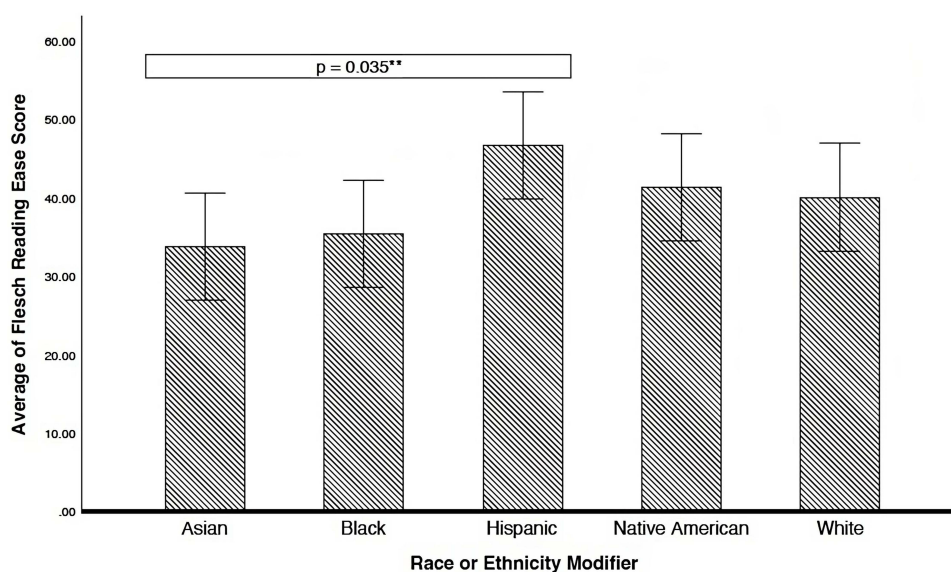


Figure 3 Flesch Reading Ease of Gemini-generated responses with race or ethnicity modifiers. Error bars represent 95% confidence intervals. ** indicates a significance value less than 0.05.

guideline, the readability of patient-oriented educational materials sourced from high-impact journals remains above recommended levels and has not significantly improved from 1998 to 2018.¹⁷

Comparing LLMs, Copilot produced materials with significantly higher difficulty levels according to Flesch-Kincaid Grade Level and Flesch-Reading Ease metrics compared to ChatGPT and Gemini. Notably, there were no significant differences observed among the platforms when assessed using the SMOG index. Previous research shows that readability formulas like Flesch-Kincaid Grade Level and Flesch Reading Ease scores may underestimate readability for materials in which comprehension is crucial. Therefore, these measures may be less suitable for health education materials compared to the SMOG index, which includes criteria for expected complete comprehension.¹³

While previous studies have shown racial, ethnic, and gender biases in material produced by machine learning models, our study suggests these biases may not always manifest consistently across platforms.¹⁸ As each algorithm is

developed on different selected materials, these can create variability in the biases expressed by different LLMs. Particularly noteworthy is the lack of detectable gender bias in the readability of materials from any platform. Prior literature identifies GPT-4 as perpetuating racial and gender stereotypes in other contexts by not appropriately modeling the demographic diversity of medical conditions.¹⁹ Our findings highlight the diverse biases inherent in different LLMs, emphasizing the need for caution when using these tools in clinical settings. Moreover, these findings underscore the need for continued efforts in training, bias detection, and further research to address the nuanced biases propagated by various language models as they continue to proliferate.

This study did not include age as a demographic variable, which may have generated more readable material based on suspected reading levels for different age groups. Beyond these findings, the LLM-generated responses varied by demographics, and future studies should examine the accuracy of responses. This study is also limited by the inherent unpredictability (stochasticity) of LLMs, and results may vary depending on when the LLM is prompted. All queries were conducted on the same day, and multiple trials were performed to reduce confounding factors. Another limitation of this study is inherent in the readability formulas used, which do not account for the imaging and videos included in the generated responses. This study followed a narrow line of inquiry about myopia, and the type of medical questions patients would ask remains unknown. This work could be replicated in a variety of contexts. Niche and rare diagnoses could further affect the readability of the generated material if information without medical jargon is less available.

Conclusion

In summary, race and ethnicity modifiers but not gender affected the length and readability of patient education materials about myopia generated by Gemini. ChatGPT and Copilot demonstrated no significant differences in readability based on demographic prompts. The breadth and accuracy of the information included based on LLM and demographic variables should be further investigated. Future studies could assess the accuracy of generated information and the readability of other disease processes to identify potential sources of misinformation. The potential impact of readability on LLM responses when prompting for certain ages could also be examined.

Abbreviations

ANOVA, analysis of variance; LLM, large language model; SMOG, Simple Measure of Gobbledygook.

Funding

This work was supported by funding from the NIH Center Core Grant P30EY014801, Research to Prevent Blindness—Unrestricted Grant (GR004596).

Disclosure

The author reports no conflicts of interest in this work. This paper/the abstract of this paper was presented at the Association for Research in Vision and Ophthalmology 2024 Conference as a poster presentation with interim findings. The poster's abstract was published in "ARVO Annual Meeting Abstracts" in *Investigative Ophthalmology & Visual Science* June 2024, Vol.65, 352: Hyperlink (<https://iovs.arvojournals.org/article.aspx?articleid=2796258>).

References

1. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med*. 2023;29(8):1930–1940. doi:10.1038/s41591-023-02448-8
2. Alowais SA, Alghamdi SS, Alsuhbany N, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC Med Educ*. 2023;23(1):689. doi:10.1186/s12909-023-04698-z
3. Ott MA. Bias in, bias out: ethical considerations for the application of machine learning in pediatrics. *J Pediatr*. 2022;247:124. doi:10.1016/j.jpeds.2022.01.035
4. Sveen W, Dewan M, Dexheimer JW. The risk of coding racism into pediatric sepsis care: the necessity of antiracism in machine learning. *J Pediatr*. 2022;247:129–132. doi:10.1016/j.jpeds.2022.04.024

5. Tyson A, Pasquini G, Spencer A, Funk C. 60% of Americans would be uncomfortable with provider relying on ai in their own health care. Available from: <https://www.pewresearch.org/science/2023/02/22/60-of-americans-would-be-uncomfortable-with-provider-relying-on-ai-in-their-own-health-care/>. Accessed January 29, 2024.
6. Schulz PJ, Nakamoto K. Patient behavior and the benefits of artificial intelligence: the perils of “dangerous” literacy and illusory patient empowerment. *Patient Educ Couns*. 2013;92(2):223–228. doi:10.1016/j.pec.2013.05.002
7. University of Miami. Institutional review boards [homepage on the Internet]. Available from: <https://hsro.uresearch.miami.edu/institutional-review-boards/index.html>. Accessed October 14, 2024.
8. QuickFacts: United States [homepage on the Internet]. United States Census Bureau; Available from: <https://www.census.gov/quickfacts/>. Accessed February 14, 2024.
9. Openai platform. [homepage on the Internet]. OpenAI. Available from: <https://platform.openai.com/docs/models/gpt-3-5-turbo>. Accessed February 18, 2024.
10. Brown D, Data SD, Privacy, and security for Microsoft copilot for Microsoft 365. [homepage on the Internet]. Microsoft. Available from: <https://learn.microsoft.com/en-us/microsoft-365-copilot/microsoft-365-copilot-privacy>. Accessed February 18, 2024.
11. Welcome to copilot in windows - Microsoft support. [Homepage on the Internet]. Microsoft. Available from: <https://support.microsoft.com/en-us/windows/welcome-to-copilot-in-windows-675708af-8c16-4675-afeb-85a5a476ccb0>. Accessed February 29, 2024.
12. Mc Laughlin GH. SMOG grading-a new readability formula. *J Read*. 1969;12(8):639–646.
13. Wang LW, Miller MJ, Schmitt MR, Wen FK. Assessing readability formula differences with written health information materials: application, results, and recommendations. *Res Social Adm Pharm*. 2013;9(5):503–516. doi:10.1016/j.sapharm.2012.05.009
14. Kao G, Thompson JS. Racial and ethnic stratification in educational achievement and attainment. *Annu Rev Sociol*. 2003;29(1):417–442. doi:10.1146/annurev.soc.29.010202.100019
15. Serxner S. How readability of patient materials affects. *J Vasc Nurs*. 2000;18(3):97–101. doi:10.1067/mvn.2000.109281
16. Weiss BD Health literacy: a manual for clinicians. [Homepage on the Internet]. American Medical Association Foundation; 2003. Available from: <http://lib.ncfh.org/pdfs/6617.pdf>. Accessed February 29, 2024.
17. Rooney MK, Santiago G, Perni S, et al. Readability of patient education materials from high-impact medical journals: a 20-year analysis. *J Patient Exp*. 2021;8:2374373521998847. doi:10.1177/2374373521998847
18. Huang J, Galal G, Etemadi M, Vaidyanathan M. Evaluation and mitigation of racial bias in clinical machine learning models: scoping review. *JMIR Med Inform*. 2022;10(5):e36388. doi:10.2196/36388
19. Zack T, Lehman E, Suzgun M, et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. *Lancet Digit Health*. 2024;6(1):e12–22. doi:10.1016/S2589-7500(23)00225-X

Clinical Ophthalmology

Dovepress

Publish your work in this journal

Clinical Ophthalmology is an international, peer-reviewed journal covering all subspecialties within ophthalmology. Key topics include: Optometry; Visual science; Pharmacology and drug therapy in eye diseases; Basic Sciences; Primary and Secondary eye care; Patient Safety and Quality of Care Improvements. This journal is indexed on PubMed Central and CAS, and is the official journal of The Society of Clinical Ophthalmology (SCO). The manuscript management system is completely online and includes a very quick and fair peer-review system, which is all easy to use. Visit <http://www.dovepress.com/testimonials.php> to read real quotes from published authors.

Submit your manuscript here: <https://www.dovepress.com/clinical-ophthalmology-journal>