Ophthalmol Ther (2025) 14:1281–1295
https://doi.org/10.1007/s40123-025-01142-x

ORIGINAL RESEARCH

Large Language Models: Pioneering New Educational Frontiers in Childhood Myopia

Mohammad Delsoz · Amr Hassan · Amin Nabavi · Amir Rahdar · Brian Fowler · Natalie C. Kerr · Lauren Claire Ditta · Mary E. Hoehn · Margaret M. DeAngelis · Andrzej Grzybowski · Yih-Chung Tham · Siamak Yousefi

Received: February 12, 2025 / Accepted: April 2, 2025 / Published online: April 21, 2025

ABSTRACT

Introduction: This study aimed to evaluate the performance of three large language models (LLMs), namely ChatGPT-3.5, ChatGPT-4o (o1 Preview), and Google Gemini, in producing patient education materials (PEMs) and improving the readability of online PEMs on childhood myopia.

Methods: LLM-generated responses were assessed using three prompts. Prompt A requested to "Write educational material on childhood myopia." Prompt B added a modifier specifying "a sixth-grade reading level using the FKGL (Flesch-Kincaid Grade Level) readability formula." Prompt C aimed to rewrite existing PEMs to a sixth-grade level using FKGL. Reponses were assessed for quality (DISCERN tool), readability (FKGL, SMOG (Simple Measure of Gobbledygook)), Patient Education Materials

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s40123-025-01142-x.

M. Delsoz · A. Nabavi · A. Rahdar · B. Fowler · N. C. Kerr · L. C. Ditta · M. E. Hoehn · S. Yousefi (✉)
Hamilton Eye Institute, Department of Ophthalmology, University of Tennessee Health Science Center, 930 Madison Ave., Suite 471, Memphis, TN 38163, USA
e-mail: Siamak.Yousefi@uthsc.edu

A. Hassan
Department of Ophthalmology, Gavin Herbert Eye Institute, University of California, Irvine, CA, USA

M. M. DeAngelis
Departement of Ophthalmology, University at Buffalo, Buffalo, NY, USA

M. M. DeAngelis
Research Service, VA Western New York Healthcare System, Buffalo, NY, USA

A. Grzybowski
Institute for Research in Ophthalmology, Foundation for Ophthalmology Development, Poznan, Poland

Y.-C. Tham
Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Republic of Singapore

Y.-C. Tham
Centre of Innovation and Precision Eye Health, Department of Ophthalmology, Yong Loo Lin School of Medicine, National University of Singapore and National University Health System, Singapore, Republic of Singapore

S. Yousefi
Department of Genetics, Genomics, and Informatics, University of Tennessee Health Science Center, Memphis, TN, USA

Assessment Tool (PEMAT, understandability/actionability), and accuracy.

Results: ChatGPT-4o (01) and ChatGPT-3.5 generated good-quality PEMs (DISCERN 52.8 and 52.7, respectively); however, quality declined from prompt A to prompt B ($p = 0.001$ and $p = 0.013$). Google Gemini produced fair-quality (DISCERN 43) but improved with prompt B ($p = 0.02$). All PEMs exceeded the 70% PEMAT understandability threshold but failed the 70% actionability threshold (40%). No misinformation was identified. Readability improved with prompt B; ChatGPT-4o (01) and ChatGPT-3.5 achieved a sixth-grade level or below (FGKL 6 ± 0.6 and 6.2 ± 0.3), while Google Gemini did not (FGKL 7 ± 0.6). ChatGPT-4o (01) outperformed Google Gemini in readability ($p < 0.001$) but was comparable to ChatGPT-3.5 ($p = 0.846$). Prompt C improved readability across all LLMs, with ChatGPT-4o (o1 Preview) showing the most significant gains (FKGL 5.8 ± 1.5; $p < 0.001$).

Conclusions: ChatGPT-4o (o1 Preview) demonstrates potential in producing accurate, good-quality, understandable PEMs, and in improving online PEMs on childhood myopia.

Keywords: Large language models; Patient education materials; Childhood myopia

Key Summary Points

This study evaluates the ability of artificial intelligence (AI)-powered tools, including ChatGPT-3.5, ChatGPT-4o (o1 Preview), and Google Gemini, to generate and improve the readability of existing educational materials on childhood myopia.

It demonstrates that these models, particularly ChatGPT-4o (o1 Preview), can produce good-quality, understandable, accurate, user-friendly content that meets readability standards, and improves the readability of existing online educational materials on childhood myopia at or below the sixth-grade level.

The study also highlights the importance of prompt customization in enhancing content clarity and provides practical recommendations for leveraging AI tools to improve health education.

The study identifies areas for improvement, such as enhancing actionability and incorporating multimedia elements.

INTRODUCTION

Childhood myopia, commonly referred to as nearsightedness, has emerged as a significant global public health concern. Recent studies indicate that approximately one-third of children and adolescents worldwide are affected by myopia, with projections estimating that the global prevalence will reach nearly 40% by 2050, equating to over 740 million cases in this age group [1]. The prevalence among children is particularly alarming; studies have reported that in East Asian countries, up to 80–90% of young adults are myopic [2]. In the USA, the prevalence of myopia in children has increased substantially over recent decades [3].

The increasing incidence of myopia in children has been related to various factors, including environmental influences, and genetic predisposition [4]. Notably, the COVID-19 pandemic and associated lockdowns have exacerbated this issue by increasing screen time and reducing outdoor activities, both of which are significant risk factors for the development and progression of myopia [1]. This shift in behavior has led to a notable rise in myopia cases among children during and after the pandemic period.

Effective patient education materials (PEMs) are crucial in managing and mitigating this condition, as they have the ability to enhance societal understanding of myopia and encourage proactive health behaviors. Effective management and mitigation of childhood myopia heavily rely on comprehensive PEMs [5]. These resources are vital for enhancing understanding and promoting proactive health behaviors among patients and their families. However, existing PEMs often exceed the sixth-grade reading level as recommended by American Medical Association (AMA), rendering them less accessible to the general population [6–8].

Recently, the advent of large language models (LLMs), including OpenAI's ChatGPT, Google's Gemini, and other artificial intelligence (AI) chatbots, has introduced innovative applications in medicine, particularly in ophthalmology. Recent studies have demonstrated promising results showing that their proficiency in natural language processing enables them to interpret complex medical data, assist in diagnosing ocular conditions, helping in research, and provide personalized treatment recommendations [9–15]. These AI-driven models have demonstrated potential in health information sector across various medical fields. Yet, their effectiveness in creating PEMs specifically tailored for childhood myopia and improving the readability of the existing PEMs remains underexplored.

This study purposes to assess the quality, readability, actionability, and accuracy of PEMs on childhood myopia generated by LLMs, including ChatGPT-3.5, ChatGPT-4o (01 Preview version), and Google Gemini. By assessing these models' capabilities in producing understandable and actionable health information and enhancing the readability of existing online resources, we seek to determine their viability as supplementary tools in patient education.

METHODS

The study was exempt from ethical review of The University of Tennessee Health Science Center as it did not involve human participants or their personal data, focusing instead on evaluating the performance of the latest AI models. The focus on publicly available data and AI-generated text ensured compliance with privacy and research ethics standards. The study took place from October to December 2024, following the principles of the Declaration of Helsinki.

Study Design

This study aimed to assess how useful LLMs are in creating PEMs for childhood myopia and improving the readability of current online resources. It compared the performance of ChatGPT-3.5, ChatGPT-4o (o1 Preview), and Gemini in generating new PEMs and rewriting existing ones to make them more accessible (Supplemental-1 in the Supplementary Material).

Large Language Models Selections

The LLMs evaluated in this study were ChatGPT-3.5, ChatGPT-4o (o1 Preview), and Gemini. These models were selected for their widespread accessibility and demonstrated capabilities in generating and refining text. ChatGPT models were accessed via the OpenAI platform, and Gemini was accessed through its dedicated web interface.

Prompt Design

To generate PEMs on childhood myopia, two distinct prompts were crafted:

- Prompt A (control): Write educational material on childhood myopia that the average American can easily understand. This general prompt was deliberately crafted without specific readability guidelines to establish a baseline understanding of the inherent readability and quality of LLM-generated PEMs.
- Prompt B (modified): "Since the average American can read at a sixth-grade reading level, utilizing the FKGL (Flesch-Kincaid Grade Level) readability formula, write educational material on childhood myopia that is easy for the average American to understand?" (as advised by the American Medical Association) (Supplemental-2 in the Supplementary Material). This prompt explicitly instructed the models to generate materials suitable for a sixth-grade reading level, referencing FKGL readability formula directly. We chose this detailed wording to assess how effectively LLMs could produce targeted,

readable content when provided with explicit guidelines, as recommended by AMA standards for PEM readability.

In prompt B, we referenced the Flesch-Kincaid Grade readability formula when creating patient educational materials for patient-targeted level by each LLM primarily because it is widely recognized and endorsed by prominent organizations, including the AMA and the National Institutes of Health (NIH) [16, 17], for assessing PEMs. Unlike other readability metrics, FKGL specifically evaluates sentence length and word complexity, two critical components that affect comprehension, making it especially suited for health information targeted toward a broad audience. It is a valuable tool, especially in educational and publishing fields, to ensure that information is clear and accessible to the intended audience [17].

Each LLM was tasked with responding to these prompts (A and B) in 20 trials, ensuring separate and unbiased chat instances for each trial to avoid the influence of Reinforcement Learning from Human Feedback (RLHF) features of the LLM chatbots [18, 19].

For improving already available online resources, first 20 eligible existing online PEMs on childhood myopia were sourced from the first two pages of a Google search engine using the keyword “childhood myopia” from October to December 2024. We acknowledge the concern that restricting our search to the first 20 eligible Google search results may introduce selection bias, as these listings might not represent all existing online resources. However, our rationale for using this approach is rooted in real-world user behavior. Numerous studies indicate that most individuals searching for health information online seldom venture beyond the first two pages of search results, as they account for more than 95% of total web traffic among search engine users [20, 21]. By focusing on the materials patients are most likely to encounter, we aimed to evaluate the educational content that exerts the greatest practical impact on health literacy and patient education. While we recognize this does not capture every possible resource, it does reflect the sources patients are statistically more likely to see. Excluded materials included advertisements, academic articles, chapters from books, multimedia content such as videos, personal weblogs, websites in languages other than English, resources for clinical decision-making, and non-patient-targeted sources. Each PEM was inputted into the LLMs with the following prompt:

Prompt C: “Since patient education materials (PEMs) are recommended to be written at a sixth-grade reading level, could you rewrite the following text to meet this standard using the FKGL readability formula? [insert text]?” Prompt C is designed to evaluate the LLMs’ ability to revise existing materials; this prompt instructed the models to rewrite existing PEMs specifically to meet a sixth-grade reading level using FKGL. The explicit reference to FKGL again allowed us to assess the models’ proficiency in adjusting existing content complexity to patient-target level, reflecting real-world scenarios where existing PEMs must be adapted for improved readability. Finally, our approach involved collecting 20 PEMs per LLM (60 in total) in prompt A, generating another 20 per LLM in prompt B (60 more), and evaluating 20 existing online educational resources and improving 20 existing online educational resources per three LLMs (60 total). This resulted in the analysis of 200 educational materials handouts overall. We used online readabilityformulas.com, a free online tool for calculating readability scores and other key metrics [22]. This software employs proven readability tools for its calculations and has been widely used in notable medical studies [23–25]. Essential readability metrics, such as syllable count, word count, complex word count, sentence count, SMOG (Simple Measure of Gobbledygook) readability score, and FKGL score, were assessed for all PEMs.

Evaluation Metrics of the Generated PEMs

We utilized the full 16-item validated DISCERN instrument for the quality and reliability of the generated PEMs and rewritten PEMs for evaluations [26]. This evaluation tool includes a 16-item questionnaire designed for use by reviewers. Each item is rated on a scale from 1 to 5, where scores of 1–2 denote low quality, 3 represents moderate quality, and 4–5 indicate high quality [27]. The DISCERN questionnaire is organized into three sections. The first section (questions 1–8) focuses on the reliability of the publication. It evaluates aspects such as the publication's purpose, the credibility of its information sources, its relevance, and whether it offers additional resources about the condition. The second section (questions 9–15) examines the details provided about treatments, including how they work, their potential risks and benefits, and whether alternative treatment options are presented. Finally, the third section (question 16) assesses the overall quality of the publication as a resource for treatment information (Supplemental-3A in the Supplementary Material). We calculated the total DISCERN score by summing all 16 item ratings (range 16–80) for each reviewer per educational material. We then mapped the DISCERN score to the corresponding quality level (Table 1), following the recommended calculation methods and analyzed it. A grading system, outlined in Table 1, was developed based on the scores earned by each PEMs [28].

Two evaluators (ophthalmologists) assessed the generated and rewritten PEM of 120 handouts. To minimize bias, evaluators were blinded to each other's assessments, the identities of the sources generating the responses were removed, and the final score for each PEM was determined by taking the median score from the reviewers. Additionally, evaluators assessed the understandability and actionability of the PEMs using the Patient Education Materials Assessment Tool (PEMAT), a validated instrument from the Agency for Healthcare Research and Quality (AHRQ) [29, 30] (Supplemental-3B in the Supplementary Material). Understandability, defined as the ability of diverse audiences to comprehend and explain core messages, was calculated as an overall percentage based on 12 yes/no questions covering items such as clarity and word choice. Actionability, referring to the ease with which patients could identify actionable next steps, was measured using five targeted yes/no questions. Materials scoring 70% or higher were classified as "understandable" or "actionable" [31, 32]. Beyond this threshold, PEMAT scores provided a basis for comparative evaluations. Lastly, the accuracy of the generated PEMs was evaluated using a Likert scale for misinformation, ranging from 1 to 5. A score of 1 meant "no misinformation," 3 indicated "moderate misinformation," and 5 represented "high misinformation" [27]. This thorough assessment provided a well-rounded analysis of the quality, usability, and reliability of educational materials on childhood myopia.

Table 1 DISCERN score grading and its quality equivalent levels

DISCERN score	Out of 100 (%)	Quality level
64–80	80 and above	Excellent
52–63	65–79	Good
41–51	51–64	Fair
30–40	37–50	Poor
16–29	20–36	Very poor

Readability Assessment of PEMs

To evaluate how easy it is to read and understand all the PEMs, we used two well-known readability metrics—the SMOG and the FKGL formulas—via the online readability calculator Readable.com. These metrics assess the grade level required to comprehend the material, with SMOG focusing on estimates of how many years of education someone would need to understand a given piece of writing. SMOG focuses on the number of polysyllabic words within a set number of sentences, providing a grade-level score [33]. The FKGL metric determines readability by looking at the average number of syllables in each word and the number of words in each sentence, yielding a US school grade level [34]. Scores ranging from 1 through

12 reflect reading difficulty levels that correspond to grades 1 through 12, scores between 13 and 16 indicate text complexity at the college level, and any score of 17 or higher signifies that a text demands a more advanced understanding than a college education typically provides [35, 36]. Formulas below represent details of these two metrics:

$$\begin{aligned}\text{Flesch} - \text{Kincaid Grade Level} \\ = 0.39\left(\text{total words}/\text{total sentences}\right) \\ + 11.8\left(\text{total syllables}/\text{total words}\right)\ 15.59\end{aligned}$$

$$\text{SMOG grade level} = 1.0430 \times \sqrt{\left(\text{number of polysyllabic words} \times \left(30 \div \text{number of sentences}\right)\right)} + 3.1291$$

Statistical Analysis

We used two-sample *t* tests to compare readability metrics and Mann–Whitney *U* tests for quality, understandability, actionability, and accuracy. We then compared the readability scores across all three language models using a one-way analysis of variance (ANOVA). To identify any significant differences in performance between the models, we followed up with Tukey's honestly significant difference (HSD) test. A significance level of $p<0.05$ was set for all analyses. Statistical analyses were conducted using SPSS (version 29, IBM Corp, USA) software.

RESULTS

All generated PEMs in prompt A by ChatGPT-4o (01-Preview version) and ChatGPT-3.5 were of good quality (median DISCERN score of 52.8 and 52.7, respectively). The quality of PEMs generated by ChatGPT-4o (o1 Preview) and ChatGPT-3.5 showed a significant decline from prompt A to prompt B, respectively ($p=0.001$ and $p=0.013$). The PEMs generated by Google Gemini were of fair quality with median DISCERN score of 43. However, the quality of PEMs increased from prompt A to prompt B ($p=0.02$). All responses generated by the three LLMs exceeded the 70% threshold for being classified as "understandable" (based on the PEMAT understandability scale, which ranges from 0 to 100%, with scores of 70% or higher considered "understandable") (Table 2).

Therefore, based on the PAMET understandability scale, all the generated PEMs had understandable details and information, clear purpose and include only the most important information, avoiding any distracting details, sufficient layout and design, logical sequence, with headers to separate sections and a summary of key points. None of the responses generated by the LLMs met the criteria to be classified as "actionable" (as defined by the PEMAT actionability scale, where scores range from 0 to 100%, with 70% or higher considered actionable). Across all 120 responses, the LLMs consistently scored 40%. This indicates that the content and structure of the generated educational materials did not clearly provide step-by-step guidance for patients to take actionable steps.

All 120 responses received a score of 1 on the Likert misinformation scale, indicating that these three LLMs did not produce any misinformation across the 120 newly generated PEMs.

Our readability analysis revealed that prompt B produced more readable PEMs compared to prompt A, as reflected by lower SMOG and FKGL scores for both ChatGPT-3.5 and ChatGPT-4o (o1 Preview) ($p<0.001$). This improvement was reflected in key readability metrics, such as a reduction in syllables, word count, 3+ syllable words (complex words), and sentence count ($p<0.05$) (Table 3).

Among these LLMs, both ChatGPT-4o (o1 Preview) and ChatGPT-3.5 could generate educational materials at or below the sixth-grade reading level in response to prompt B, respectively (FGKL scores 6 ± 0.6; FGKL scores 6.2 ± 0.3), while Google Gemini could not produce material at this grade level (FGKL scores; 7 ± 0.6).

In head-to-head analysis, ChatGPT-4o (o1 Preview) generated PEMs more readable (lower SMOG and FGKL scores; 5.8 ± 0.7 and 6 ± 0.6, respectively) than the Google Gemini ($p<0.001$), although the difference was not statistically

Table 2 Comparing prompt A and prompt B: evaluating the quality and understandability of large language models in generated patient education materials

LLMs	Discern points	Median (range)	*N*	Mean ranks	Sum ranks	*p* value*
Quality (DISCERN)						
Gemini	Prompt A	43 (43–43)	20	16.52	330	0.021
	Prompt B	49 (37–57)	20	24.51	490	
ChatGPT3.5	Prompt A	52.72 (51–53)	20	24.18	483.52	0.013
	Prompt B	51 (49–53)	20	16.83	336.54	
ChatGP-4o (01 Preview)	Prompt A	52.81 (49–53)	20	25.38	507.53	0.001
	Prompt B	50 (41–53)	20	15.63	312.52	
Understandability (PEMAT %)						
Gemini	Prompt A	75 (75–83.3)%	20	20.53	410	1.00
	Prompt B	75 (75–83.3)%	20	20.52	410	
ChatGPT3.5	Prompt A	75 (75–83.3)%	20	19	380	0.389
	Prompt B	83.31 (75–83.3)%	20	22	440	
ChatGP-4o (01 Preview)	Prompt A	75 (75–83.3)%	20	20	400	0.771
	Prompt B	83.32 (75–83.3)%	20	21	420	

LLM large language model, *PEMAT* Patient Education Materials Assessment Tool

*Mann–Whitney *U* test conducted between prompt A and B (significance at $p < 0.05$)

significant compared to ChatGPT-3.5 ($p = 0.846$) (Fig. 1).

Using prompt C, the three LLMs significantly improved the readability of existing online educational resources, as evidenced by a marked enhancement in average readability scores ($p \leq 0.001$) (Fig. 2). The original online PEMs about the childhood myopia had a mean SMOG score of 10.3 ± 2.2 and an FKGL score of 9.7 ± 1.9 which is well above the recommended sixth-grade reading level (Table 4).

After applying the LLMs, significant improvements were observed ($p < 0.001$). ChatGPT-3.5 reduced the readability scores to a SMOG of 7.6 ± 1.2 and an FKGL of 7.7 ± 1.4. Google Gemini also contributed to improved readability, with a SMOG of 7.8 ± 1.3 and an FKGL of 7.5 ± 1.1, while ChatGPT-4o (o1 Preview) demonstrated the most significant enhancement, achieving or staying below the specified sixth-grade reading level (SMOG 5.3 ± 1.6, FKGL 5.8 ± 1.5). Additionally, in a head-to-head analysis of readability improvements made by the LLMs, ChatGPT-4o (o1 Preview) consistently outperformed both ChatGPT-3.5 and Google Gemini by a significant margin ($p < 0.001$ for both comparisons) (Fig. 3). These results highlight the transformative potential of LLMs, especifically ChatGPT-4o (o1 Preview), in simplifying complex medical materials, making them more accessible and user-friendly for patients.

DISCUSSION

To our knowledge, this is the first study to explore how LLMs can assist parents in understanding online health information and generating PEMs about childhood myopia. Unlike earlier studies that primarily relied on standardized exams and related queries for evaluation [37–39], our research takes a different approach by exploring realistic scenarios where worried

Table 3 Performance of LLMs based on readability metrics on prompt A vs prompt B

Readability metrics	ChatGPT-3.5			ChatGPT-4o (01 preview)			Google Gemini		
	Prompt A	Prompt B	*p* value (A vs B)	Prompt A	Prompt B	*p* value (A vs B)	Prompt A	Prompt B	*p* value (A vs B)
Syllables	789.7 (137.5)	562.4 (76.9)	< 0.001	861.9 (145.8)	534.4 (80)	< 0.001	558.9 (109.2)	407.5 (107.8)	< 0.001
Words	500.2 (96.9)	381.6 (51.2)	< 0.001	564.5 (91.8)	370.5 (51.6)	< 0.001	326.6 (60.2)	272.2 (60.7)	< 0.001
3+ syllable words	37.3 (6.3)	18.4 (4.1)	< 0.001	36.7 (10.8)	15.35 (5)	< 0.001	35.2 (10.9)	17.6 (7.3)	< 0.001
Sentences	27.5 (5.6)	27.3 (3.7)	< 0.001	31.7 (4.9)	24.5 (4.3)	< 0.001	19.7 (4.2)	17.6 (4.9)	< 0.001
SMOG Readability Score	8.03 (0.6)	5.9 (0.4)	< 0.001	7.7 (0.7)	5.8 (0.7)	< 0.001	9.07 (0.62)	6.7 (0.8)	< 0.001
Flesch-Kincaid Grade Level	7.7 (0.5)	6.2 (0.3)	< 0.001	7.5 (0.58)	6 (0.6)	< 0.001	8.5 (0.7)	7.0 (0.6)	< 0.001

LLM large language model, *SMOG* Simple Measure of Gobbledygook

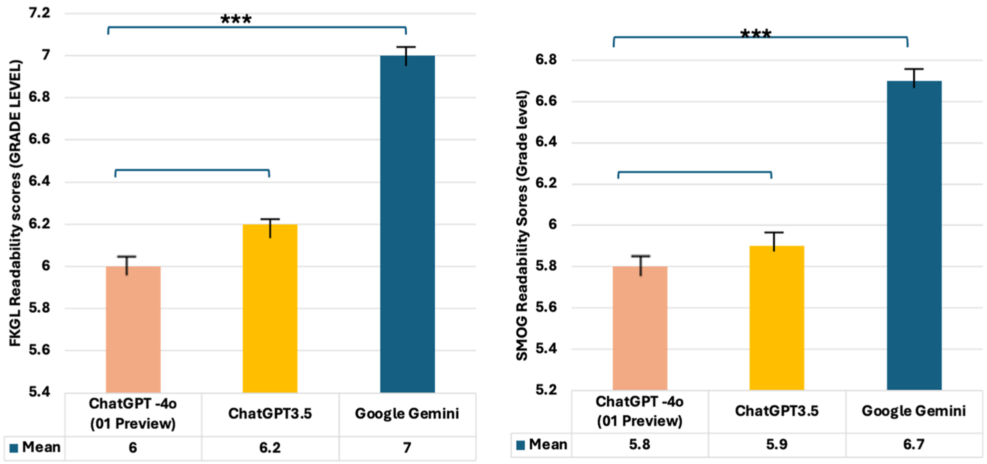

Fig. 1 Comparing the performance of large language models for prompt B based on SMOG (Simple Measure of Gobbledygook), and FKGL (Flesch-Kincaid Grade Level) scores. One-way ANOVA (one-way analysis of variance), post hoc Tukey test

parents might turn to these emerging tools for help. This highlights the critical need to assess how accurate and reliable the responses from LLM chatbots are in real-world situations. Therefore, we assessed the performance of LLMs, namely ChatGPT-4o (01Preview), ChatGPT-3.5, and Google Gemini, using a comprehensive approach that included DISCERN scores to evaluate the quality of the materials, PEMAT scores to measure their understandability and actionability, a misinformation scale to check for accuracy, and readability metrics to ensure

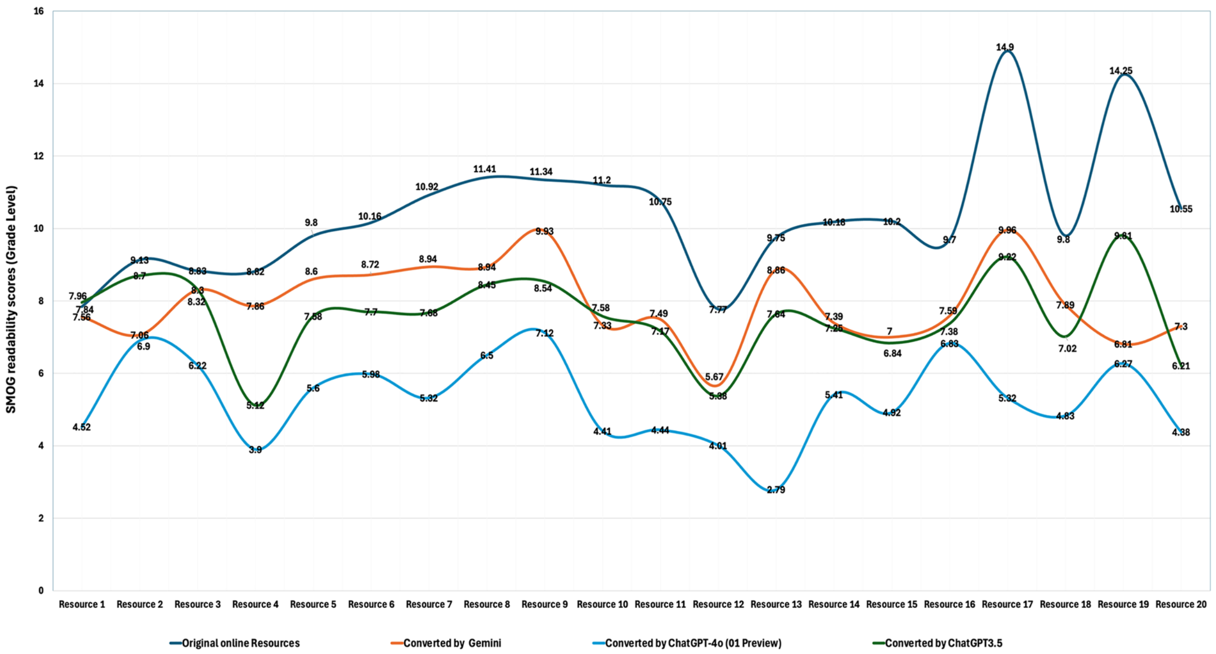

Fig. 2 Readability of Online educational resources and performance of ChatGPT-4o (01 Preview), ChatGPT-3.5, and Google Gemini in improving the readability of online original resources. *SMOG* Simple Measure of Gobbledygook

Table 4 Readability of original online resources and performance of LLMs for improving the readability of the original online resources

Readability metrics	Original resources	ChatGPT-4o (01 preview)	*p* value (Orig. vs GPT-4o)	ChatGPT-3.5	*p* value (Orig. vs GPT-3.5)	Google Gemini	*p* value (Orig. vs Gemini)
Syllables	1457.2 (736.9)	330.4 (159.6)	< 0.001	984.5 (412.7)	0.008	808.4 (205.7)	< 0.001
Words	871.0 (392.7)	230.4 (107.9)	< 0.001	649.2 (256.2)	0.21	521.9 (122.4)	< 0.001
3+ syllable words	91.0 (60.7)	13.3 (10.4)	< 0.001	43.0 (23.1)	0.001	38.5 (12.6)	< 0.001
Sentences	41.5 (17.1)	18.4 (7.8)	0.04	41.0 (16.6)	0.9	31.1 (9.0)	0.02
SMOG Readability Score	10.3 (2.2)	5.3 (1.6)	< 0.001	7.6 (1.2)	< 0.001	7.8 (1.3)	< 0.001
Flesch-Kincaid Grade Level	9.7 (1.9)	5.8 (1.5)	< 0.001	7.7 (1.4)	< 0.001	7.5 (1.1)	< 0.001

LLM large language model, *SMOG* Simple Measure of Gobbledygook

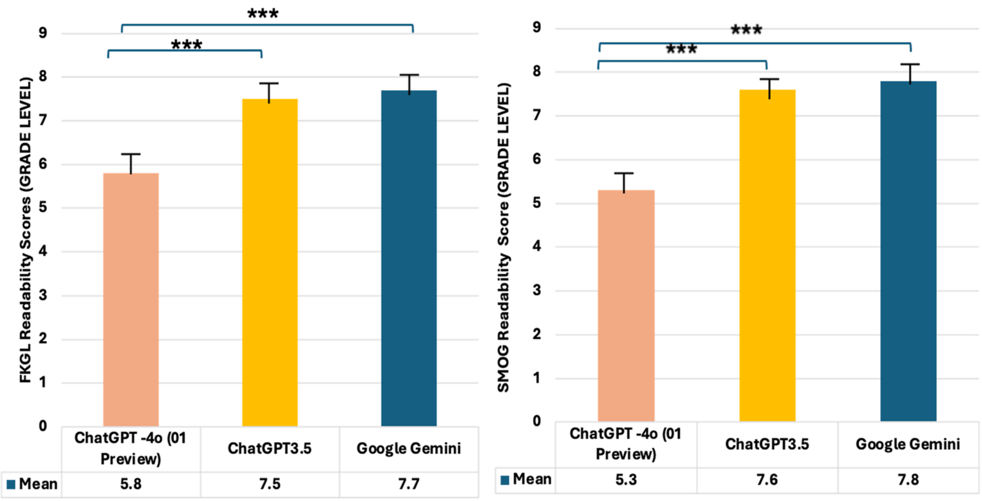

Fig. 3 Performance of large language models for rewriting original handouts based on SMOG (Simple Measure of Gobbledygook), and FKGL (Flesch-Kincaid Grade Level) scores. One-way ANOVA (one-way analysis of variance), post hoc Tukey test

the content was easy to comprehend. All generated PEMs in prompt A by ChatGPT-4o (o1 Preview) and ChatGPT-3.5 were of good quality, while the PEMs generated by Google Gemini were of fair quality. However, this should not be understood as a clear superiority as ChatGPT over Google Gemini. Previous studies found that each AI model has its own strengths and weaknesses with Google Gemini being superior in aspects like language understanding but lacking in other aspects in which ChatGPT excels [40]. All responses generated by the three LLMs were understandable but none were actionable, as in they did not clearly provide step-by-step guidance for patients to take actionable steps. We identified shortcomings in two key areas of actionability (PEMAT items 23 and 26), highlighting opportunities for improvement. First, the responses lacked practical tools such as action reminder checklists or planners to help patients follow through with the information. Second, since the responses were entirely text-based, they missed the added clarity and engagement that instructional visual aids could provide. It is worth considering that incorporating patient action tools into the phrasing of prompts could be a promising area for future research to enhance the actionability of upcoming PEMs. This finding is very interesting because despite AI showing promise in developing management plans for patients in previous studies [41], our paper shows that it is unreliable in generating action plans for patients aiming to organize their checklists when dealing with childhood myopia. This is most likely not an inherent weakness in AI itself, but rather in the datasets on which it was trained [41].

Although health educational materials must be understandable and easily readable, they must also have high-quality information. There was not any misinformation across the 120 newly generated PEMs. Our findings highlight that the LLMs evaluated in this study did not produce any AI "confabulations" or generate false or misleading information, a common issue reported in other studies [42, 43].

After evaluations of the top 20 search results, as they account for more than 95% of total web traffic among users of online web search engines [20], all results exceed the sixth-grade reading level recommended by the NIH and the AMA, which is suggested for ensuring the average American can understand health information [6, 8, 44]. All the evaluated LLMs demonstrated the ability to rewrite or improve the readability of online educational resources. However, ChatGPT-4o (o1 Preview) outperformed other models, including ChatGPT-3.5 and Google Gemini, by significantly enhancing readability and consistently meeting or staying below the grade level recommended by organizations like the NIH and AMA for the average American to understand health information [16, 44]. ChatGPT-4o (o1 Preview) adapted the content to a reading level at or below the average person's proficiency, making the information more accessible and easier to understand for a broader audience. While ChatGPT-3.5 and Google Gemini improved the readability of online PEMs, they did not consistently meet the recommended standards set by the NIH and AMA for understanding health information. This underscores ChatGPT-4o (o1 Preview)'s superior ability to simplify complex medical language into more digestible formats for users.

ChatGPT-4o (o1 Preview) has shown superiority in generating the most readable LLMs in other diseases as well. In one study, ChatGPT-4o (o1 Preview) was tested for its usefulness in teaching individuals about dermatological conditions, and researchers found it could create paragraphs understandable by those with a reading level ranging from high school to early college [45]. Another evaluation explored ChatGPT-4o (o1 Preview)'s ability to produce information on uveitis by prompting it to write patient-targeted health details about this condition. Appropriateness and readability were then assessed with the latter measured using the FKGL formula; in all attempts, the content proved both suitable and easily adjustable [46]. In another highly relevant study, researchers evaluated the performance of LLM chatbots for producing patient-targeted health materials on childhood glaucoma using a prompting method for gauging text comprehensibility. ChatGPT-4o (o1 Preview) demonstrated promising results [7]. This could be because ChatGPT-4o (o1 Preview) was trained on a larger dataset of written materials and has more parameters than its earlier versions [47]. However, as of now, ChatGPT-4o (o1 Preview) is only available through a paid subscription, which may make it less accessible to the general public. Our findings reveal that the way a prompt is structured can significantly impact the readability of the responses generated by an LLM. All three models created more readable content when provided with a more detailed query (prompt B) compared to a less specific one (prompt A). Prompt B was more detailed, as it included a modifier statement that specified the target grade level for readability and explicitly instructed the use of the "FKGL readability formula." Our findings highlight the critical importance of using validated tools, such as readability formulas, to achieve the desired level of readability in educational materials. This approach aligns with recent studies, emphasizing that incorporating such tools ensures content is both accessible and effective for its intended audience [7, 46].

This study highlights the immense potential of LLMs, particularly ChatGPT-4o (o1 Preview), as a tool for bridging gaps in health literacy and providing tailored, patient-centered educational materials. However, the lack of actionability in the generated content underscores the need for further advancements in AI model development to create step-by-step, actionable health guidance for users. Future research should focus on exploring methods to optimize LLM outputs, such as integrating domain-specific datasets and refining prompt engineering techniques to elicit more actionable responses. Additionally, studies incorporating diverse linguistic and cultural contexts would help generalize these findings beyond English-speaking populations and address global health literacy challenges. Given the increasing reliance on visuals in health communication, future studies should also evaluate the ability of LLMs to generate or recommend high-quality visual aids, such as infographics and interactive tools, to complement textual materials. Such advancements could facilitate the integration of LLMs into clinical workflows, enabling healthcare providers to create more personalized and comprehensible patient education resources in real time.

In summary, these findings highlight the ability of LLMs to serve as a versatile tool for delivering clear, producing good quality of PEMs, enhancing the readability of online educational resources, accessible health information and this could open the door for integrating LLM chatbots into the management of myopia care after more advancement and refinements.

Our study has some limitations. As LLMs are not perfect, the responses generated by LLMs are also limited to the data on which they are trained. Moreover, the way prompts are phrased is inherently subjective. To minimize potential bias, we carefully broke down the prompts into distinct control and modifier sub-statements and conducted multiple repeated trials for each prompt while also relying on established principles of prompt engineering [47]. Furthermore, we only evaluated PEM in the English language, which, although the most commonly used, but is not always the specific language in which patients would be interested in reading about their medical conditions. Moreover, for evaluating existing online resources, we recognize that limiting our scope to the top 20 eligible PEMs in Google has the potential to exclude lesser-ranked sites that could also provide valuable or alternative information. However, capturing the most commonly accessed PEMs reflects a real-world usage pattern and therefore remains a practical and meaningful sample. Additionally, this study primarily focuses on the US population, as most of the research on readability and health literacy comes from studies conducted in the USA. That said, it is important to acknowledge that inadequate health literacy is a widespread issue in many other major countries as well [48, 49]. Finally, many patient-focused health resources incorporate visuals, such as graphics, images, videos, and demonstrations, to help improve understanding. However, our study did not evaluate these features.

CONCLUSIONS

This study highlights the promising potential of large language models, particularly ChatGPT-4o (o1 Preview), which, at the time of this investigation, had the best performance of the evaluated LLMs in generating good-quality, readable, and understandable patient education materials for childhood myopia. While ChatGPT-4o (o1 Preview) outperformed other models in enhancing readability and accessibility, a significant limitation was the lack of actionable guidance in the generated content. This highlights the need for continued improvements in AI models to develop more actionable health information. Future research should focus on refining prompt engineering, incorporating diverse linguistic and cultural contexts, and exploring multimodal features like visuals to fully realize the potential of LLMs in patient education. With continued improvements, LLMs could play a key role in making healthcare information more accessible, personalized, and effective.

Medical Writing/Editorial Assistance. During the preparation of this manuscript, the authors utilized ChatGPT-4 from OpenAI for editing and proofreading purposes to enhance readability. Subsequently, the authors carefully reviewed and refined the content as necessary and assume full responsibility for the final version of the publication.

Author Contributions. Mohammad Delsoz, Amr Hassan, Amin Nabavi, Amir Rahdar, Brian Fowler, Natalie C. Kerr, Lauren Claire Ditta, Mary E. Hoehn, Margaret M DeAngelis, Andrzej Grzybowski, Yih-Chung Tham, and Siamak Yousefi contributed to conception of the study. Mohammad Delsoz, Amr. Hassan, Brian Fowler, Lauren Claire Ditta, Margaret M DeAngelis, Andrzej Grzybowski, Yih-Chung Tham, and Siamak Yousefi contributed to study design. Mohammad Delsoz, Amr. Hassan, Amin Nabavi, Amir Rahdar, and Siamak Yousefi contributed to acquisition of the data. Mohammad Delsoz and Siamak Yousefi contributed to contributed to analysis, interpretation of the data, figures, and tables. Mohammad Delsoz, Amr. Hassan, and Siamak Yousefi accessed and verified each dataset during the study. The research planning and execution were supervised by Siamak Yousefi, Yih-Chung Tham, Margaret M DeAngelis, Andrzej Grzybowski, Brian Fowler, and Lauren Claire Ditta, Mohammad Delsoz, Amr Hassan, Lauren Claire Ditta, Andrzej Grzybowski, Yih-Chung Tham, and Siamak Yousefi drafted the manuscript. All authors read and approved the final version of the manuscript.

Funding. This work was supported by Hamilton Eye Institute (SY). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The journal's Rapid Service Fee was funded by Dr. Siamak Yousefi.

Data Availability. All essential data for replicating our results is in the Supplementary file, except raw grader scores, which are available upon request.

Declarations

Conflict of Interest. Mohammad Delsoz, Amr Hassan, Amin Nabavi, Amir Rahdar, Brian Fowler, Natalie C. Kerr, Lauren Claire Ditta, Mary E. Hoehn, Margaret M DeAngelis, and Yih-Chung Tham have nothing to disclose. Andrzej Grzybowski is an Editorial Board member of *Ophthalmology and Therapy*. Andrzej Grzybowski was not involved in the selection of peer reviewers for the manuscript nor any of the subsequent editorial decisions. Siamak Yousefi: Received prototype instruments from Remidio, M&S Technologies, and Visrtucal Fields. He gives consultations to the InsihgtAEye and Enolink.

Ethical Approval. The study was exempt from ethical review of The University of Tennessee Health Science Center as it did not involve human participants or their personal data, focusing instead on evaluating the performance of the latest AI models. The focus on publicly available data and AI-generated text ensured compliance with privacy and research ethics standards. The study took place from October to December 2024, following the principles of the Declaration of Helsinki.

REFERENCES

1. Liang J, Pu Y, Chen J, et al. Global prevalence, trend and projection of myopia in children and adolescents from 1990 to 2050: a comprehensive systematic review and meta-analysis. Br J Ophthalmol. 2024. https://doi.org/10.1136/bjo-2024-325427.

2. Modjtahedi BS, Ferris FL, Hunter DG, Fong DS. Public health burden and potential interventions for myopia. Ophthalmology. 2018;125(5):628–30.

3. Schweitzer K. With nearsightedness in children on the rise, experts push for outdoor time, disease designation. JAMA. 2024;332(19):1599–601. https://doi.org/10.1001/jama.2024.21043.

4. Morgan IG, Wu P-C, Ostrin LA, et al. IMI risk factors for myopia. Investig Ophthalmol Vis Sci. 2021;62(5):3–3. https://doi.org/10.1167/iovs.62.5.3.

5. Huang J, Wen D, Wang Q, et al. Efficacy comparison of 16 interventions for myopia control in children: a network meta-analysis. Ophthalmology. 2016;123(4):697–708.

6. Stossel LM, Segar N, Gliatto P, Fallar R, Karani R. Readability of patient education materials available at the point of care. J Gen Intern Med. 2012;27:1165–70.

7. Dihan Q, Chauhan MZ, Eleiwa TK, et al. Using large language models to generate educational materials on childhood glaucoma. Am J Ophthalmol. 2024;265:28–38.

8. 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.

9. Raja H, Huang X, Delsoz M, et al. Diagnosing glaucoma based on the ocular hypertension treatment study dataset using chat generative pre-trained

transformer as a large language model. Ophthalmol Sci. 2025;5(1):100599.

10. Huang X, Raja H, Madadi Y, et al. Predicting glaucoma before onset using a large language model chatbot. Am J Ophthalmol. 2024. https://doi.org/10.1016/j.ajo.2024.06.035.

11. Delsoz M, Madadi Y, Raja H, et al. Performance of ChatGPT in diagnosis of corneal eye diseases. Cornea. 2024;43(5):664–70.

12. Delsoz M, Raja H, Madadi Y, et al. The use of ChatGPT to assist in diagnosing glaucoma based on clinical case reports. Ophthalmol Ther. 2023;12(6):3121–32.

13. Madadi Y, Delsoz M, Lao PA, et al. ChatGPT assisting diagnosis of neuro-ophthalmology diseases based on case reports. J Neuroophthalmol. 2022;128:1356.

14. Madadi Y, Delsoz M, Khouri AS, Boland M, Grzybowski A, Yousefi S. Applications of artificial intelligence-enabled robots and chatbots in ophthalmology: recent advances and future trends. Curr Opin Ophthalmol. 2024;35(3):238–43.

15. Bellanda VC, Santos MLD, Ferraz DA, Jorge R, Melo GB. Applications of ChatGPT in the diagnosis, management, education, and research of retinal diseases: a scoping review. Int J Retina Vitreous. 2024;10(1):79.

16. Weiss BD. Health literacy. Am Med Assoc. 2003;253:358.

17. Readability Scoring System. Readability formulas. https://www.readabilityformulas.com. Accessed 7 Mar 2024.

18. Delsoz M, Raja H, Madadi Y, et al. A response to: letter to the editor regarding "The use of ChatGPT to assist in diagnosing glaucoma based on clinical case reports." Ophthalmol Ther. 2024;13(6):1817–9. https://doi.org/10.1007/s40123-024-00937-8.

19. OpenAI. Introducing ChatGPT. OpenAI. 2022.

20. Insights C. The value of Google result positioning. Westborough: Chitika; 2013. p. 0–10.

21. Morahan-Martin JM. How internet users find, evaluate, and use online health information: a cross-cultural review. CyberPsychol Behav. 2004;7(5):497–510. https://doi.org/10.1089/cpb.2004.7.497.

22. Readability Scoring System. Readability formulas. https://www.readabilityformulas.com/.

23. Martin CA, Khan S, Lee R, et al. Readability and suitability of online patient education materials for glaucoma. Ophthalmol Glaucoma. 2022;5(5):525–30.

24. Decker H, Trang K, Ramirez J, et al. Large language model-based chatbot vs surgeon-generated informed consent documentation for common procedures. JAMA Netw Open. 2023;6(10):e2336997.

25. Crabtree L, Lee E. Assessment of the readability and quality of online patient education materials for the medical treatment of open-angle glaucoma. BMJ Open Ophthalmol. 2022;7(1):e000966.

26. Charnock D, Shepperd S, Needham G, Gann R. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. J Epidemiol Community Health. 1999;53(2):105–11.

27. Pan A, Musheyev D, Bockelman D, Loeb S, Kabarriti AE. Assessment of artificial intelligence chatbot responses to top searched queries about cancer. JAMA Oncol. 2023;9(10):1437–40.

28. San Giorgi MRM, de Groot OSD, Dikkers FG. Quality and readability assessment of websites related to recurrent respiratory papillomatosis. Laryngoscope. 2017;127(10):2293–7. https://doi.org/10.1002/lary.26521.

29. Shoemaker SJ, Wolf MS, Brach C. The patient education materials assessment tool (PEMAT) and user's guide. Rockville: Agency for Healthcare Research and Quality; 2020.

30. Shoemaker SJ, Wolf MS, Brach C. Development of the Patient Education Materials Assessment Tool (PEMAT): a new measure of understandability and actionability for print and audiovisual patient information. Patient Educ Couns. 2014;96(3):395–403.

31. Veeramani A, Johnson AR, Lee BT, Dowlatshahi AS. Readability, understandability, usability, and cultural sensitivity of online patient educational materials (PEMS) for lower extremity reconstruction: a cross-sectional study. Plast Surg. 2024;32(3):452–9.

32. Loeb S, Sengupta S, Butaney M, et al. Dissemination of misinformative and biased information about prostate cancer on YouTube. Eur Urol. 2019;75(4):564–7.

33. Edmunds MR, Barry RJ, Denniston AK. Readability assessment of online ophthalmic patient information. JAMA Ophthalmol. 2013;131(12):1610–6. https://doi.org/10.1001/jamaophthalmol.2013.5521.

34. Lois C, Edward L. Assessment of the readability and quality of online patient education materials for the medical treatment of open-angle glaucoma. BMJ Open Ophthalmol. 2022;7(1):e000966. https://doi.org/10.1136/bmjophth-2021-000966.

35. Kincaid P, Fishburne RP, Rogers RL, Chissom BS. Derivation of new readability formulas (automated readability index, fog count and Flesch reading ease formula) for Navy Enlisted Personnel. Millington: Naval Air Station Memphis; 1975.

36. Mc Laughlin GH. SMOG grading—a new readability formula. J Read. 1969;12(8):639–46.

37. Antaki F, Touma S, Milad D, El-Khoury J, Duval R. Evaluating the performance of ChatGPT in ophthalmology: an analysis of its successes and shortcomings. Ophthalmol Sci. 2023;3(4):100324.

38. Mihalache A, Popovic MM, Muni RH. Performance of an artificial intelligence chatbot in ophthalmic knowledge assessment. JAMA Ophthalmol. 2023;141(6):589–97.

39. Lim ZW, Pushpanathan K, Yew SME, et al. Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. EBioMedicine. 2023;95:104770.

40. Masalkhi M, Ong J, Waisberg E, Lee AG. Google DeepMind's Gemini AI versus ChatGPT: a comparative analysis in ophthalmology. Eye. 2024;38(8):1412–7. https://doi.org/10.1038/s41433-024-02958-w.

41. Satapathy SK, Kunam A, Rashme R, Sudarsanam PP, Gupta A, Kumar HK. AI-assisted treatment planning for dental implant placement: clinical vs AI-generated plans. J Pharm Bioallied Sci. 2024;16(Suppl 1):S939.

42. Hua H-U, Kaakour A-H, Rachitskaya A, Srivastava S, Sharma S, Mammo DA. Evaluation and comparison of ophthalmic scientific abstracts and references by current artificial intelligence chatbots. JAMA Ophthalmol. 2023;141(9):819–24. https://doi.org/10.1001/jamaophthalmol.2023.3119.

43. Brender TD. Medicine in the era of artificial intelligence: hey chatbot, write me an H&P. JAMA Intern Med. 2023;183(6):507–8. https://doi.org/10.1001/jamainternmed.2023.1832.

44. Doak LG, Doak CC, Meade CD. Strategies to improve cancer education materials. Oncol Nurs Forum. 1996;23:1305–12.

45. Mondal H, Mondal S, Podder I. Using ChatGPT for writing articles for patients' education for dermatological diseases: a pilot study. Indian Dermatol Online J. 2023;14(4):482–6.

46. Kianian R, Sun D, Crowell EL, Tsui E. The use of large language models to generate education materials about uveitis. Ophthalmol Retina. 2024;8(2):195–201.

47. Brin D, Sorin V, Vaid A, et al. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. Sci Rep. 2023;13(1):16492.

48. DeWalt DA, Berkman ND, Sheridan S, Lohr KN, Pignone MP. Literacy and health outcomes: a systematic review of the literature. J Gen Intern Med. 2004;19:1228–39.

49. Schillinger D. Social determinants, health literacy, and disparities: intersections and controversies. HLRP Health Literacy Res Pract. 2021;5(3):e234–43.