

Leveraging Large Language Models to Improve Readability of Online Patient Education Materials: Cross-sectional Study

John Will, Mahin Gupta, Jonah Zaretsky, Aliesha Dowlath, Paul Testa, Jonah Feldman

Submitted to: Journal of Medical Internet Research
on: December 11, 2024

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript..... 5

Supplementary Files..... 29

 Figures 30

 Figure 1..... 31

 Multimedia Appendixes 32

 Multimedia Appendix 0..... 33

 Multimedia Appendix 0..... 33

Leveraging Large Language Models to Improve Readability of Online Patient Education Materials: Cross-sectional Study

John Will^{1*} MPA; Mahin Gupta^{1*}; Jonah Zaretsky^{2*} MD; Aliesha Dowlath^{1*} RN, MN; Paul Testa^{1, 3*} MD, JD, MPH; Jonah Feldman^{1, 4*} MD

¹ MCIT Department of Health Informatics New York University Langone Health New York US

² Division of Hospital Medicine Department of Medicine New York University Langone Health New York US

³ Ronald O. Perelman Department of Emergency Medicine NYU Grossman School of Medicine New York US

⁴ Department of Medicine NYU Long Island School of Medicine Mineola US

*these authors contributed equally

Corresponding Author:

John Will MPA

MCIT Department of Health Informatics
New York University Langone Health
New York
US

Abstract

Background: Online accessible patient education materials (PEMs) are essential for patient empowerment. However, studies have shown that these materials often exceed the recommended sixth grade reading level, making them difficult for many patients to understand. Large language models (LLMs) have the potential to transform PEMs into more readable educational content.

Objective: We sought to evaluate whether three LLMs—ChatGPT, Gemini, and Claude—can optimize the readability of PEMs to the recommended reading level without compromising accuracy.

Methods: This cross-sectional study used 60 randomly selected PEMs available online from three websites. We prompted LLMs to simplify the reading level of online PEMs. The primary outcome was the readability of the original online PEMs compared with the LLM-simplified versions. Readability scores were calculated using four validated indices Flesch Reading Ease (FRE), Flesch-Kincaid Grade Level (FKGL), Gunning Fog Index (GFI), and Simple Measure of Gobbledygook Index (SMOGI). Accuracy and understandability were also assessed as balancing measures, with understandability measured using the Patient Education Materials Assessment Tool-Understandability (PEMAT-U).

Results: The original readability scores for the American Heart Association (AHA), American Cancer Society (ACS), and American Stroke Association (ASA) websites were above the recommended sixth-grade level, with mean grade level scores of 10.7, 10.0, and 9.6, respectively. After optimization by the LLMs, readability scores significantly improved across all three websites when compared to the original text. Compared to the original website, Wilcoxon Signed-Rank Test showed ChatGPT improved the readability to 7.6 from 10.1 ($P < .001$), Gemini improved to 6.6 ($P < .001$), and Claude improved to 5.6 ($P < .001$). Word counts were significantly reduced by all LLMs, with a decrease from a mean of 410.9 to 953.9 words to a mean range of 201.9 to 248.1 words. None of the ChatGPT LLM-simplified PEMs were inaccurate, while 3.3% of Gemini and Claude LLM-simplified PEMs were inaccurate. Baseline understandability scores, as measured by PEMAT-U, were preserved across all LLM-simplified versions.

Conclusions: This cross-sectional study demonstrates that LLMs have the potential to significantly enhance the readability of online PEMs, making them more accessible to a broader audience. However, variability in model performance and demonstrated inaccuracies underscore the need for human review of LLM output. Further study is needed to identify the most reliable and effective LLM based approaches to improve readability.

(JMIR Preprints 11/12/2024:69955)

DOI: <https://doi.org/10.2196/preprints.69955>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to all users.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in [JMIR Publications](#), my full manuscript will be available to all users.

No. Please do not make my accepted manuscript PDF available to anyone.

Original Manuscript

Leveraging Large Language Models to Improve Readability of Online Patient Education Materials: Cross-sectional Study

Abstract

Background: Online accessible patient education materials (PEMs) are essential for patient empowerment. However, studies have shown that these materials often exceed the recommended sixth grade reading level, making them difficult for many patients to understand. Large language models (LLMs) have the potential to simplify PEMs into more readable educational content.

Objective: We sought to evaluate whether three LLMs—ChatGPT, Gemini, and Claude—can optimize the readability of PEMs to the recommended reading level without compromising accuracy.

Methods: This cross-sectional study used 60 randomly selected PEMs available online from three websites. We prompted LLMs to simplify the reading level of online PEMs. The primary outcome was the readability of the original online PEMs compared with the LLM-simplified versions. Readability scores were calculated using four validated indices Flesch Reading Ease (FRE), Flesch-Kincaid Grade Level (FKGL), Gunning Fog Index (GFI), and Simple Measure of Gobbledygook Index (SMOGI). Accuracy and understandability were also assessed as balancing measures, with understandability measured using the Patient Education Materials Assessment Tool-Understandability (PEMAT-U).

Results: The original readability scores for the American Heart Association (AHA), American Cancer Society (ACS), and American Stroke Association (ASA) websites were above the recommended sixth-grade level, with mean grade level scores of 10.7, 10.0, and 9.6, respectively. After optimization by the LLMs, readability scores significantly improved across all three websites when compared to the original text. Compared to the original website, Wilcoxon Signed-Rank Test showed ChatGPT improved the readability to 7.6 from 10.1 ($P<.001$), Gemini improved to 6.6 ($P<.001$), and Claude improved to 5.6 ($P<.001$). Word counts were significantly reduced by all LLMs, with a decrease from a mean range of 410.9 to 953.9 words to a mean range of 201.9 to 248.1

words. None of the ChatGPT LLM-simplified PEMs were inaccurate, while 3.3% of Gemini and Claude LLM-simplified PEMs were inaccurate. Baseline understandability scores, as measured by PEMAT-U, were preserved across all LLM-simplified versions.

Conclusions: This cross-sectional study demonstrates that LLMs have the potential to significantly enhance the readability of online PEMs while maintaining accuracy and understandability, making them more accessible to a broader audience. However, variability in model performance and demonstrated inaccuracies underscore the need for human review of LLM output. Further study is needed to explore advanced LLM techniques and models trained for medical content.

Keywords: patient education; health literacy; artificial intelligence

Introduction

Online health information is increasingly becoming a leading source for which United States (US) adults seek health advice. More individuals use online health information for health advice than family, friends, coworkers, health care professionals, or traditional media [1]. Health information sourced from online materials can supplement the patient-physician conversation regarding health advice [2], however many individuals without access to a healthcare provider only seek information online [3]. This is concerning as online materials are not tailored to an individual patient's specific needs and require readers to discern appropriately between valuable and less relevant health information [4]. Given that 36% of US adults have only basic or below basic health literacy skills [5], navigating the abundant health information online becomes particularly challenging.

Health literacy is an individual's ability to understand and use health information to make informed decisions [6]. The inability to interpret health information can have serious consequences for patients, as individuals with low health literacy rates often experience poorer health outcomes [7]. They may also be more likely to be readmitted to the emergency department or access inappropriate health services than those with adequate health literacy, contributing to rising healthcare costs [8,9]. There is a growing recognition that the creators of health information must present materials in a simplified format that encourages understanding, as making patient education materials (PEMs) more readable and understandable has been shown to lead to better comprehension of health-related information among those with low health literacy [6,10]. Several validated measures of readability exist, including the Flesch Reading Ease [11] (FRE), Flesch-Kincaid Grade Level [12] (FKGL), Gunning Fog Index [13] (GFI), and the Simple Measure of Gobbledygook Index [14] (SMOGI). Additionally, understandability can be measured by the Patient Education Materials Assessment Tool-Understandability [15] (PEMAT-U) scores.

The National Institutes of Health (NIH) recommends that PEM is written at a maximum reading level of sixth grade [16]. However, PEMs provided by hospitals and those available online are

consistently above a sixth grade reading level [17-19]. Further, one study reviewing 100 PEMs from three distinct PEM content vendors providing materials to the National Library of Medicine and electronic health records vendors found most of the PEMs were above an eighth grade reading level [20]. A metanalysis of 7,891 websites using 13 different readability scales found mean grade level scores ranged from grade 10 to 15 [21]. Disease specific studies focusing on stroke and heart failure found materials are consistently above the recommended reading level [22-24], which is especially concerning as they are among the leading causes of death in the US [25]. While improvements in readability scores have been made over time [17] and human intervention can assist in improving readability [19], challenges remain in bringing readability to an appropriate level on a mass scale.

Large language models (LLMs) can consume massive amounts of data and perform a wide array of tasks that have made them useful tools to understand and generate text [26], making them an enticing option to improve the readability of PEMs. One study evaluated the impact of institutionally developed PEMs using LLMs and found that readability could be improved, but still did not meet the NIH recommended reading level [27]. In another study seeking to improve patient accessibility of published discharge summaries, LLMs enhanced readability and reliably reached the targeted reading level. However, the study concluded that LLM outputs required clinician oversight to ensure accuracy [28].

There are multiple publicly accessible LLMs available for use today, including ChatGPT [29], Gemini [30], and Claude [31]. Potentially, patients or PEM creators could use these resources to simplify online PEM material, but little is known about the efficacy or safety of this approach, though risks of bias and transparency have been identified [32]. Existing literature on LLM simplification often focuses on institution-specific or topic-specific materials [33-35], which limits the broader applicability of the findings. These studies typically address readability improvements within a narrow context, leaving a gap in a wholistic understanding of the impact on readability, understandability, and accuracy [27,33,36]. There is a need for research that evaluates PEMs across a

wider range of health topics and sources to provide more generalizable insights. As such, the primary objective of this study was to evaluate if LLMs, when prompted, could simplify readability of online PEM content across a range of health topics and sources, while maintaining accuracy and understandability.

Methods

This cross-sectional study looked at online accessible PEMs accessed between 7/1/24 and 8/30/24. All study procedures complied with institutional ethical standards and those set by the Declaration of Helsinki and are reported using both the Strengthening the Reporting of Observational Studies in Epidemiology [37] (STROBE) reporting guidelines for cross-sectional studies and the TRIPOD-LLM checklist for reporting studies involving the use of LLMs [38]. The project does not involve human subjects research and did not require institutional review.

Recruitment

The study used 60 randomly selected PEMs available online from three websites. The number 60 was chosen a priori based on feasibility. We selected three websites with publicly available health education materials, the American Heart Association [39] (AHA), American Cancer Society [40] (ACS), and the American Stroke Association [41] (ASA). We selected these three organizations as heart disease, cancer, and stroke are among the leading causes of death in the US [25]. Materials for review were selected from the organization's website. Twenty articles from each site were randomly selected for review.

Intervention

We chose to use free publicly available generative artificial intelligence (AI) platforms to improve the readability of the 60 PEMs, and we accessed the platforms through the publicly available chat interfaces. We chose free platforms as they help to support equitable access to the technology for all patients; for this reason, we intentionally accessed the platforms through the same web-based interface patients experience. These platforms were OpenAI's ChatGPT, Google's Gemini, and

Anthropic's Claude. Each of the 60 PEMs was entered into the freely available LLMs ChatGPT (Version: GPT-4, released on May 13, 2024, OpenAI, San Francisco, CA, USA), Gemini (Version: Gemini-1.5-flash, released on May 24, 2024, Google, Mountain View, CA, USA), and Claude (Version: Claude 3.5 Sonnet, released on June 20, 2024, Anthropic, San Francisco, CA, USA).

We asked each LLM via prompt to "Translate to a fifth grade reading level" and pasted the original article text into the LLM. We selected this prompt to model research showing that AI tools improved readability by targeting a lower level than the sixth grade level recommended by the NIH. This approach, which has been proven successful, was designed to account for variability in grade level interpretations by LLMs [27]. We then saved copies of each output from the LLM, resulting in four total versions of each material (the original plus three LLM versions).

Measures

The primary outcome of readability of each PEM and the LLM-simplified PEM was measured by online accessible scoring of FRE, FKGL, GFI, and SMOGI [42]. The PEMAT-U score was manually scored by a project team member using the PEMAT guide [15]. We also extracted the number of words for each PEM and each LLM-simplified PEM. We completed assessments for errors or inaccuracies by reviewing the LLM-simplified PEM and comparing it to the original PEM as the criterion standard. Two project team members in non-clinical roles completed these reviews independently. If either team member marked the LLM-simplified PEM as having errors or inaccuracies, a physician member of the project team reviewed the article to determine the final outcome. Errors or inaccuracies were defined as instances in which the original message was changed, or the LLM simplified content misrepresented the meaning of the original content that could lead to a different interpretation by the reader.

Statistical Analysis

Descriptive statistics presented include means and standard deviation of the FRE, FKGL, GFI, SMOGI, and PEMAT-U scores, as well as the number of words to measure material length

independently of the score. We used a One-Way ANOVA test to test for significant differences in the original PEM of the material between websites. To assess statistically significant differences in readability scores and word count from the original PEM to the LLM-simplified PEM, we employed Wilcoxon Signed-Rank Test to compare each LLM-simplified PEM to its corresponding original PEM. All data was prepared and analyzed in IBM SPSS Version 28.0.1.1 [43].

Results

We evaluated 60 unique PEMs from three different websites [eTable1]. Mean readability scores across the three websites were not significantly different from each other, regardless of the readability employed. The AHA had the highest FKGL, GFI, and SMOGI reading scores, and the lowest FRE score, though none were significantly different. The ACS had significantly more words than both the AHA and ASA.

All three LLMs had significantly improved readability scores when compared to the original source. Gemini and Claude had greater improvements in readability than ChatGPT for each website. Though improved, there were only two instances in which the project target of achieving a mean sixth grade reading level was met (Chat GPT's AHA FKGL score and Gemini's ACA FKGL and GFI score and ASA GFI score). There were five instances in which the LLM prompt of achieving a mean fifth grade reading level was met (ChatGPT's ASA FKGL score, Gemini's AHA GFI score, Claude's AHA GFI score, and Claude's ACS and ASA GFI score). In five other instances, the mean reading level was improved beyond a fifth grade reading level (Gemini's AHA and ASA FKGL score, and Claude's AHA, ACS, and ASA FKGL score). Readability scores and standard deviations for each website's original PEM and the three LLM-simplified PEM are shown in Table 1.

Table 1. Mean reading scores and standard deviations of the original version of the patient education material and each large language model version (N = 60).

Website	Score	Original	ChatGPT	P-value	Gemini	P-value	Claude	P-value
American Heart	FRE	52.9 (11.0)	74.2 (7.4)	<.001	86.3(7.8)	<.001	83.6 (6.3)	<.001

Association	FKGL	9.1 (1.7)	6.1 (1.3)	<.001	3.9 (1.3)	<.001	3.8 (0.9)	<.001
	GFI	11.3 (2.0)	8.0 (1.3)	<.001	5.8 (1.1)	<.001	5.6 (1.0)	<.001
	SMOGI	11.7 (1.4)	9.1 (1.1)	<.001	7.3 (1.0)	<.001	7.3 (0.8)	<.001
	PEMAT-(U %)	75.5 (13.1)	74.7 (11.4)	.776	81.1 (8.5)	.149	85.4 (5.8)	.011
	# of Words	663.0 ^a (286.6)	213.3 (54.1)	<.001	213.3 (54.1)	<.001	243.5 (65.6)	<.001
American Cancer Society	FRE	59.7 (9.0)	65.8 (10.3)	<.001	67.1 (12.3)	.009	79.4 (9.2)	<.001
	FKGL	8.5 (1.5)	7.1 (1.5)	<.001	6.6 (1.8)	<.001	4.5 (1.3)	<.001
	GFI	10.2 (1.6)	8.1 (1.5)	<.001	7.7 (2.0)	<.001	5.6 (1.0)	<.001
	SMOGI	11.1 (1.3)	9.8 (1.3)	<.001	9.4 (1.4)	<.001	7.5 (1.1)	<.001
	PEMAT-(U %)	81.3 (11.1)	76.5 (10.1)	.044	87.6 (5.8)	.029	81.6 (6.8)	.881
	# of Words	953.9 ^a (725.5)	243.6 (102.4)	<.001	243.6 (102.4)	<.001	248.1 (60.5)	<.001
American Stroke Association	FRE	59.0 (14.6)	75.0 (9.2)	<.001	80.5 (9.5)	<.001	85.6 (6.0)	<.001
	FKGL	7.9 (2.3)	5.6 (1.5)	<.001	4.2 (1.4)	<.001	3.6 (1.3)	<.001
	GFI	10.0 (2.6)	7.4 (1.7)	<.001	6.0 (1.3)	<.001	5.6 (1.5)	<.001
	SMOGI	10.8 (2.0)	8.7 (1.7)	<.001	7.8 (1.2)	<.001	7.1 (1.4)	<.001
	PEMAT-(U%)	77.5 (9.3)	75.9 (8.3)	.256	82.8 (4.9)	.046	78.9 (6.6)	.722
	# of Words	410.9 ^a (207.4)	201.9 (75.9)	<.001	201.9 (75.9)	<.001	230.8 (66.1)	<.001

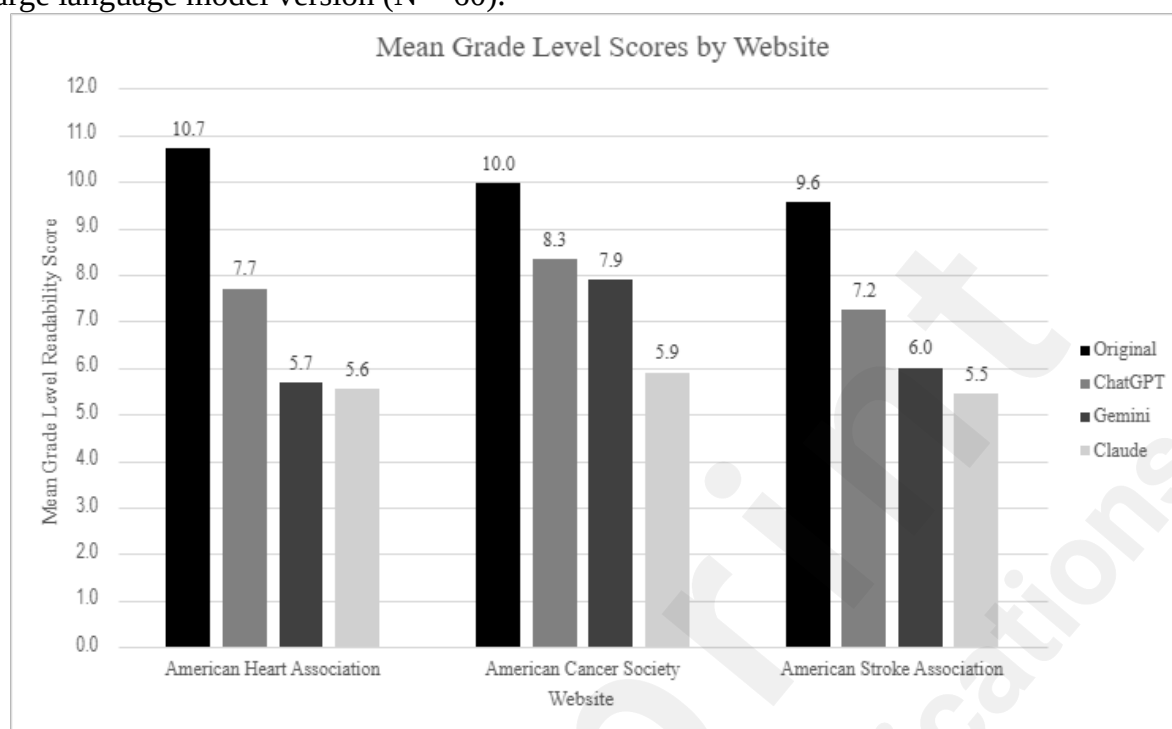
^aResults of One-Way ANOVA test indicate original sites are significantly different from each other at $p < .05$

P-values reported are the large language model text compared to original source text, using Wilcoxon Signed-Rank Test
FRE Flesch Reading Ease, *FKGL* Flesch-Kincaid Grade Level, *GFI* Gunning-Fog Index, *SMOGI* Simple Measure of Gobbledygook Index, *PEMAT-U* Patient Education Materials Assessment Tools - Understandability

We calculated the average of the three grade level readability scores (FKGL, GFI, and SMOGI) for each PEM version of the material, and compared them across sites. Compared to the original website, ChatGPT improved the readability to 7.6 from 10.1 ($P < .001$), Gemini improved to 6.6 ($P < .001$), and Claude improved to 5.6 ($P < .001$). Of LLM-simplified versions of the PEM, only Claude consistently achieved a mean fifth grade reading level on all three websites (5.5 to 5.9), while Gemini did on one of the websites, and ChatGPT on none. Figure 1 shows mean readability scores

by website and PEM version.

Figure 1. Mean grade level scores for the original version of the patient education material and each large language model version (N = 60).



We assessed accuracy for all LLM-simplified PEM content based on its original version. ChatGPT had no inaccuracies on all three sites, while Gemini experienced two inaccuracies on the AHA website and Claude experienced one inaccuracy on the ACA and one on the ASA website. Accuracy data is presented in Table 2. The four inaccuracies identified are presented in Table 3, with the original and LLM-simplified text.

Table 2. Number of inaccuracies and inaccuracy rate of each large language model (LLM) version when compared to the original, by LLM.

	ChatGPT	Gemini	Claude
American Heart Association	0.0/20.0 (0.0%)	2.0/20.0 (10.0%)	0.0/20.0 (0.0%)
American Cancer Society	0.0/20.0 (0.0%)	0.0/20.0 (0.0%)	1.0/20.0 (5.0%)
American Stroke Association	0.0/20.0 (0.0%)	0.0/20.0 (0.0%)	1.0/20.0 (5.0%)
Total	0.0/60.0 (0.0%)	2.0/60.0 (3.3%)	2.0/60.0 (3.3%)

Table 3. Examples of inaccuracies in the large language model version when compared to the

original content.

Patient Education Material Site	Original Text	Large Language Model Text	Large Language Model
American Heart Association	Coronary Microvascular Disease	Tiny Heart Trouble	Gemini
American Heart Association	Mitral Valve Stenosis	A Sticky Valve	Gemini
American Cancer Society	Solitary papillomas (solitary intraductal papillomas) are single tumors that often grow in the large milk ducts near the nipple. They are a common cause of clear or bloody nipple discharge, especially when it comes from only one breast. They may be felt as a small lump behind or next to the nipple. Sometimes they can be painful.	Sometimes, small bumps can grow inside the tiny pipes in your breast that carry milk. These bumps are not dangerous and won't hurt you. They're called intraductal papillomas.	Claude
American Stroke Association	Bathroom Transfers: Practice transfers with your therapist first so you're confident and safe.	Moving Safely: Practice moving with your helper first.	Claude

Discussion

Principal Results

Our study reveals that the readability of PEMs from the AHA, ACS, and ASA websites significantly exceeds the NIH-recommended sixth grade reading level. This finding aligns with existing literature, which consistently reports that online PEMs are often written at a higher reading level [21-24]. The mean readability scores for the original PEMs were 10.7, 10.0, and 9.6 for AHA, ACS, and ASA, respectively.

The application of LLMs such as ChatGPT, Gemini, and Claude demonstrated a substantial improvement in the readability of these materials. Still, despite being prompted to translate the text

to a fifth grade reading level, the LLMs often produced content that was above this target. ChatGPT had the highest mean reading level (7.2 to 8.3), followed by Gemini (5.7 to 7.9). Only Claude consistently achieved a mean fifth grade reading level (5.5 to 5.9), which is below the NIH recommended maximum level. These results suggest that while LLMs can enhance readability, achieving the desired reading level remains challenging. Alternative prompting techniques that include more description or seek a lower reading level score may yield different results. Although no specific patterns related to content or type of PEM was identified that scored higher or lower, FKGL readability scores were consistently lower than GFI or SMOGI scores, highlighting the variation of the readability scoring algorithms.

Understandability, as measured by the PEMAT-U, saw some significant improvements, though this was not consistent across all sites and LLMs. Some notable improvements within the PEMAT-U by the LLMs focused on word choice by only using medical terms to familiarize the audience and subsequently defining them, organization by breaking the material into short sections if the original material was not, and layout by adding bullet points.

Interestingly, our study found that Google's Gemini outperformed ChatGPT, contrary to previous research where ChatGPT was more effective than Google's Bard [27]. This variability is not unexpected, as newer models can exhibit different performance levels, but it does highlight the need for further research to identify the most reliable and effective LLM-based approaches for optimizing PEM readability

Accuracy is another critical factor when simplifying PEMs. Our study found high accuracy rates for LLM-simplified PEMs, with ChatGPT achieving 100% accuracy and Gemini and Claude at 96.7%. These rates are significantly higher than those reported in our previous work on simplified discharge summaries, which had a 54% accuracy rate [28]. The higher accuracy in this study may be attributed to the simpler nature of PEMs compared to the more technically worded discharge summaries. However, the tradeoff between simplification and accuracy remains a concern in this study as well,

as Gemini and Claude, the models that performed best at simplification, were the least accurate.

Beyond accuracy concerns, the use of AI in healthcare presents some inherent risks. One metanalysis found bias and transparency are two primary risks associated with AI in healthcare [32]. This is an important consideration when using LLMs to simplify PEMs as important content intended to minimize bias in PEM may be omitted for the sake of simplification. Additionally, LLMs are not obviously transparent about the source of data used in their text generation, presenting risks of misinformation and the need for clinical oversight.

Overall, this study demonstrates that while publicly available LLMs hold great promise, they may not yet be fully reliable for direct use by patients to simplify PEMs. However, based on our results, PEM creators who are content experts in their respective fields could certainly start to use LLMs to improve the readability of their own materials. This could be accomplished prior to the publication of materials online through prompt engineering and human review of the simplified PEM. To gain efficiency, content creators could prompt LLMs to create PEM rather than simplify it and add guardrails to ensure necessary content is maintained. This approach could help address the persistent issue of PEMs being written above the recommended reading level, which disproportionately affects individuals with low health literacy. Given the increasing reliance on online information as a primary source of health information, the adoption of LLMs by online PEM creators has great potential for enhancing patient understanding and improving health outcomes.

Recommendations

Future research should explore more advanced LLM engineering techniques, such as few-shot learning, where LLMs learn through examples included within the prompt to achieve targeted readability goals [28]. One could envision a library of prompts made available through a publicly accessible application designed specifically for this purpose. Additionally, integrating LLM self-evaluation, where the model checks its own output for inaccuracies, and generated knowledge prompting, which involves asking the model to generate its knowledge about a topic before

providing an answer to reduce hallucinations and improve accuracy, could further enhance the effectiveness of these models. While our study focused on free, publicly accessible LLMs to ensure equitable access, more sophisticated pay-for-play models like OpenAI's o1 model and Claude Pro may offer greater reliability and effectiveness. Additionally, LLMs trained specifically on medical data or fine-tuned for this specific task are promising avenues for future study.

While the out-of-the-box use of LLMs is not yet fully reliable for direct patient use, the significant opportunity for improved performance suggests that this capability is well within reach. With continued advancements and refinements, we are confident that LLM simplification of PEMs can become a safe and effective patient-facing tool that empowers patients to better understand their health conditions and ultimately improves health literacy and outcomes.

Limitations

This study has several limitations. First, it only assessed content from three websites, and results may vary with other sources of PEM, especially PEM on rare diseases where online PEM may be less abundant. Second, the study included only English-language PEMs; materials in other languages may yield different outcomes. Third, we used publicly available content and prompted the LLMs to translate the content to a specific reading level. Different methods, such as asking LLMs to create content from scratch, may produce different results. Additionally, the manual calculation of PEMAT-U scores may introduce subjective bias, and the focus on readability and understandability of text ignores the capability of LLMs to create diagrams, infographics, and figures that can help with patient understanding. Further, team members calculating the PEMAT-U scores and accuracy were not blinded to the LLM they were evaluating due to limited project resources, introducing reviewer bias. Finally, while the content was generally accurate, we did not evaluate whether important omissions were made during simplification. Future studies should assess the impact of content loss on the PEMs' educational value. Although independent reviews of the materials were completed by non-clinical team members to prevent medical knowledge bias in the evaluation of PEMAT-U and

accuracy, it's possible review by patients or patient caregivers would respond differently.

Conclusions

Online PEMs consistently exceed the NIH-recommended maximum sixth grade reading level, posing a challenge for individuals with low health literacy. LLMs offer a promising solution for simplifying PEMs to improve accessibility. Our study demonstrates that LLMs can significantly enhance readability while maintaining high accuracy. However, achieving the desired reading level remains challenging, and human oversight is necessary to ensure the completeness and accuracy of simplified content. Further research is needed to explore advanced LLM techniques and models specifically trained for medical content to optimize the readability and understandability of PEMs.

Data Sharing

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Author Contributions

JW conducted the data analysis. JF, MG, and JW oversaw the study design and execution. JW, MG, JF and AD collected data for the study. JW, MG, JF, AD, JZ, and PT edited the manuscript. All authors approved the final manuscript.

Author JF, who is dyslexic, would like to acknowledge the use of GPT4-o, which he accessed through Azure AI Studio, to improve the spelling, grammar, and clarity of his written contribution. All GPT-4-o suggestions were actively reviewed and revised. Authors take full responsibility for the originality and integrity of the content in the publication

Acknowledgements: The authors have none to report.

Conflicts of Interest: The authors have no conflicts of interest to disclose.

Abbreviations:

PEM – patient education material

LLM – large language model

AI – artificial intelligence

US – United States

AHA – American Heart Association

ACS – American Cancer Society

ASA – American Stroke Association

FRE - Flesch Reading Ease

FKGL - Flesch-Kincaid Grade Level

GFI - Gunning Fog Index

SMOGI - Simple Measure of Gobbledygook Index (SMOGI)

Patient Education Materials Assessment Tool-Understandability (PEMAT-U).

References

1. Jacobs W, Amuta AO, Jeon KC. Health information seeking in the digital age: An analysis of health information seeking behavior among US adults. Alvares C, ed. *Cogent Social Sciences*. 2017;3(1). doi:<https://doi.org/10.1080/23311886.2017.1302785>
2. Tan SSL, Goonawardene N. Internet Health Information Seeking and the Patient-Physician Relationship: A Systematic Review. *Journal of Medical Internet Research*. 2017;19(1):e9. doi:<https://doi.org/10.2196/jmir.5729>
3. Amante DJ, Hogan TP, Pagoto SL, English TM, Lapane KL. Access to care and use of the Internet to search for health information: results from the US National Health Interview Survey. *Journal of Medical Internet Research*. 2015;17(4):e106. doi:<https://doi.org/10.2196/jmir.4126>
4. Swire-Thompson B, Lazer D. Public Health and Online Misinformation: Challenges and Recommendations. *Annual Review of Public Health*. 2020;41(1):433-451. doi:<https://doi.org/10.1146/annurev-publhealth-040119-094127>
5. Cutilli CC, Bennett IM. Understanding the Health Literacy of America. *Orthopaedic Nursing*. 2009;28(1):27-32. doi:<https://doi.org/10.1097/01.nor.0000345852.22122.d6>
6. CDC. What Is Health Literacy? Health Literacy. Published October 8, 2024. https://www.cdc.gov/health-literacy/php/about/?CDC_AAref_Val=https://www.cdc.gov/healthliteracy/learn/index.html
7. Berkman ND, Sheridan SL, Donahue KE, Halpern DJ, Crotty K. Low Health Literacy and Health Outcomes: An Updated Systematic Review. *Annals of Internal Medicine*. 2011;155(2):97. doi:<https://doi.org/10.7326/0003-4819-155-2-201107190-00005>
8. Shahid R, Shoker M, Chu LM, Frehlick R, Ward H, Pahwa P. Impact of Low Health Literacy on Patients' Health outcomes: a Multicenter Cohort Study. *BMC Health Services Research*. 2022;22(1). doi:<https://doi.org/10.1186/s12913-022-08527-9>
9. Palumbo R. Examining the impacts of health literacy on healthcare costs. An evidence synthesis. *Health Services Management Research*. 2017;30(4):197-212.

doi:<https://doi.org/10.1177/0951484817733366>

10. Kim W, Kim I, Baltimore K, Imtiaz AS, Bhattacharya BS, Lin L. Simple contents and good readability: Improving health literacy for LEP populations. *International Journal of Medical Informatics*. 2020;141(September 2020):104230. doi:<https://doi.org/10.1016/j.ijmedinf.2020.104230>

11. Flesch R. A new readability yardstick. *Journal of Applied Psychology*. 1948;32(3):221-233. doi:<https://doi.org/10.1037/h0057532>

12. Kincaid J, Fishburne R, Rogers R, Chissom B. Derivation Of New Readability Formulas (Automated Readability Index, Fog Count And Flesch Reading Ease Formula) For Navy Enlisted Personnel. *Institute for Simulation and Training*. Published online January 1, 1975. https://stars.library.ucf.edu/istlibrary/56/?utm_source=stars.library.ucf.edu%2Fistlibrary%2F56&utm_medium=PDF&utm_campaign=PDFCoverPages

13. Mc Laughlin GH. SMOG Grading-a New Readability Formula. *Journal of Reading*. 1969;12(8):639-646. <https://www.jstor.org/stable/40011226>

14. Leonard Grabeel K, Russomanno J, Oelschlegel S, Tester E, Heidel RE. Computerized versus hand-scored health literacy tools: a comparison of Simple Measure of Gobbledygook (SMOG) and Flesch-Kincaid in printed patient education materials. *Journal of the Medical Library Association*. 2018;106(1). doi:<https://doi.org/10.5195/jmla.2018.262>

15. Agency for Healthcare Research and Quality. PEMAT for Printable Materials (PEMAT-P). www.ahrq.gov. Published November 2020. <https://www.ahrq.gov/health-literacy/patient-education/pemat-p.html>

16. Eltorai A, Ghanian S, Adams C, Born C, Daniels A. Readability of Patient Education Materials on the American Association for Surgery of Trauma Website. *Archives of Trauma Research*. 2014;3(1). doi:<https://doi.org/10.5812/atr.18161>

17. Badarudeen S, Sabharwal S. Assessing Readability of Patient Education Materials: Current Role in Orthopaedics. *Clinical Orthopaedics and Related Research®*. 2010;468(10):2572-2580.

doi:<https://doi.org/10.1007/s11999-010-1380-y>

18. Hutchinson N, Baird GL, Garg M. Examining the Reading Level of Internet Medical Information for Common Internal Medicine Diagnoses. *The American Journal of Medicine*. 2016;129(6):637-639. doi:<https://doi.org/10.1016/j.amjmed.2016.01.008>

19. Stiller C, Brandt L, Adams M, Gura N. Improving the Readability of Patient Education Materials in Physical Therapy. *Cureus*. Published online February 20, 2024. doi:<https://doi.org/10.7759/cureus.54525>

20. Stossel LM, Segar N, Gliatto P, Fallar R, Karani R. Readability of Patient Education Materials Available at the Point of Care. *Journal of General Internal Medicine*. 2012;27(9):1165-1170. doi:<https://doi.org/10.1007/s11606-012-2046-0>

21. Daraz L, Morrow AS, Ponce OJ, et al. Readability of Online Health Information: A Meta-Narrative Systematic Review. *American Journal of Medical Quality*. 2018;33(5):487-492. doi:<https://doi.org/10.1177/1062860617751639>

22. Wasir AS, Volgman AS, Jolly M. Assessing readability and comprehension of web-based patient education materials by American Heart Association (AHA) and CardioSmart online platform by American College of Cardiology (ACC): How useful are these websites for patient understanding? *American Heart Journal Plus: Cardiology Research and Practice*. 2023;32(2023):100308. doi:<https://doi.org/10.1016/j.ahjo.2023.100308>

23. Kher A, Johnson S, Griffith R. Readability Assessment of Online Patient Education Material on Congestive Heart Failure. *Advances in Preventive Medicine*. 2017;2017:1-8. doi:<https://doi.org/10.1155/2017/9780317>

24. Ahn AB, Sajal Kulhari, Karimi A, Sundararajan S, Sajatovic M. Readability of patient education material in stroke: a systematic literature review. *Topics in stroke rehabilitation*. 2023;31(4):345-360. doi:<https://doi.org/10.1080/10749357.2023.2259177>

25. Centers for Disease Control and Prevention. Leading causes of death. Centers for Disease

- Control and Prevention. Published May 2, 2024. <https://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm>
26. IBM. What Are Large Language models? | IBM. [www.ibm.com](https://www.ibm.com/topics/large-language-models). Published 2023. <https://www.ibm.com/topics/large-language-models>
27. Rouhi AD, Ghanem YK, Laman Yolchieva, et al. Can Artificial Intelligence Improve the Readability of Patient Education Materials on Aortic Stenosis? A Pilot Study. *Cardiology and therapy*. 2024;13(1):137-147. doi:<https://doi.org/10.1007/s40119-023-00347-0>
28. Zaretsky J, Kim JM, Baskharoun S, et al. Generative Artificial Intelligence to Transform Inpatient Discharge Summaries to Patient-Friendly Language and Format. *JAMA Network Open*. 2024;7(3):e240357. doi:<https://doi.org/10.1001/jamanetworkopen.2024.0357>
29. OpenAI. ChatGPT. <https://chatgpt.com/>
30. Google. Gemini. [gemini.google.com](https://gemini.google.com/app). Published 2024. <https://gemini.google.com/app>
31. Anthropic. Claude. [claude.ai](https://claude.ai/new). <https://claude.ai/new>
32. Chustecki M. Benefits and Risks of AI in Health Care: Narrative Review. *Interactive Journal of Medical Research*. 2024;13(e53616):e53616-e53616. doi:<https://doi.org/10.2196/53616>
33. Kirchner GJ, Kim RY, Weddle J, Bible JE. Can Artificial Intelligence Improve the Readability of Patient Education Materials? *Clinical Orthopaedics and Related Research*. 2023;481(11). doi:<https://doi.org/10.1097/corr.0000000000002668>
34. Moons P, Liesbet Van Bulck. Using ChatGPT and Google Bard to improve the readability of written patient information: a proof of concept. *European Journal of Cardiovascular Nursing*. 2023;23(2). doi:<https://doi.org/10.1093/eurjcn/zvad087>
35. Vallurupalli M, Shah ND, Vyas RM. Optimizing Readability of Patient-Facing Hand Surgery Education Materials Using Chat Generative Pretrained Transformer (ChatGPT) 3.5. *The Journal of Hand Surgery*. 2024;49(10):986-991. doi:<https://doi.org/10.1016/j.jhsa.2024.05.007>
36. Oliva AD, Pasick LJ, Hoffer ME, Rosow DE. Improving readability and comprehension levels of

otolaryngology patient education materials using ChatGPT. American Journal of Otolaryngology. 2024;45(6):104502. doi:<https://doi.org/10.1016/j.amjoto.2024.104502>

37. Equator network. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: guidelines for reporting observational studies | The EQUATOR Network. www.equator-network.org. Published March 6, 2023. <https://www.equator-network.org/reporting-guidelines/strobe/>

38. Gallifant J, Afshar M, Ameen S, et al. The TRIPOD-LLM Statement: A Targeted Guideline For Reporting Large Language Models Use. *medRxiv : the preprint server for health sciences*. Published online 2024:2024.07.24.24310930. doi:<https://doi.org/10.1101/2024.07.24.24310930>

39. American Heart Association. American Heart Association. www.heart.org. Published 2024. <https://www.heart.org/en/>

40. American Cancer Society. American Cancer Society. Cancer.org. Published 2024. <https://www.cancer.org/>

41. American Stroke Association. American Stroke Association | A Division of the American Heart Association. www.stroke.org. Published 2019. <https://www.stroke.org/en/>

42. Added Bytes. Readable. Readable. Published 2018. <https://readable.com/>

43. IBM. Gradpack and Faculty Packs - IBM SPSS Statistics. Ibm.com. Published 2024. Accessed December 9, 2024. https://www.ibm.com/products/spss-statistics/gradpack?adoper=257533_1&adobe_mc_sdid=SDID%3D7C356A822AEF881F-00B02C6E05257922%7CMCORGID%3DD10F27705ED7F5130A495C99%40AdobeOrg%7CTS%3D1733761825&adobe_mc_ref=https%3A%2F%2Fwww.google.com%2F

44. Chustecki M. Benefits and Risks of AI in Health Care: Narrative Review. Interactive Journal of Medical Research. 2024;13(e53616):e53616-e53616. doi:<https://doi.org/10.2196/53616>

Supplementary Table 1. Websites and articles titles of online patient education materials utilized in this study, including the number of words, readability, and understandability scores for each article (N = 60).

Website	Article Number	Article Title	Number of Words	FRE	FKGL	GFI	SMOGI	PEMAT-U
American Heart Association	1	Protect Your Heart in the Heat	727	68.7	6.5	6.5	9.8	87.5%
	2	Living with Diabetes	932	60.7	7.6	7.6	10.7	81.3%
	3	Smoking and Heart Disease in Women	539	67.9	7.4	7.4	10.5	93.8%
	4	Coronary Heart Disease - Coronary Heart Disease	337	49.1	9.6	9.6	12.5	75.0%
	5	Roles of Your Four Heart Valves	391	55.8	8.0	8.0	11.4	93.8%
	6	What is Diabetes?	848	44	10.3	10.3	13.0	81.3%
	7	Risks for Heart Valve Problems	351	51.5	10.2	10.2	12.5	66.7%
	8	Recurrent Pericarditis	990	42.7	10.3	10.3	12.0	75.0%
	9	Sleep Disorders and Atrial Fibrillation	503	41.9	11.9	11.9	14.2	87.5%
	10	Coronary Microvascular Disease (MVD)	1249	45.6	9.7	9.7	11.5	62.5%
	11	Ross Procedure	226	57	8.3	8.3	11.2	80.0%
	12	Problem: Mitral Valve Stenosis	443	48	10.0	10.0	12.7	66.7%
	13	The flu vaccine: What you need to know this year	314	60.5	8.7	8.7	10.9	84.6%
	14	Understanding Your Risks for Excessive Blood Clotting	647	55.4	8.5	8.5	11.8	76.9%
	15	Preventing and Managing Falls in Adults with Cardiovascular Disease	662	36.2	11.6	11.6	13.2	56.3%
	16	What is Cardiovascular Disease?	898	62.2	6.9	6.9	9.9	66.7%
	17	Illegal Drugs and Heart Disease	589	41.3	10.6	10.6	12.7	69.2%
	18	Syncope (Fainting)	1121	34.4	12.1	12.1	14.4	43.8%
	19	Peripheral	879	68.9	6.7	6.7	10.1	69.2%

		Angiography						
	20	Causes and Prevention of Heart Disease for Women	614	66.8	7.1	7.1	9.8	92.3%
American Cancer Society	21	Risk Factors and Causes of Childhood Cancer	595	62.8	8.2	8.2	10.5	85.0%
	22	Limitations of Mammograms	753	55	9.3	9.3	12.1	83.0%
	23	Intraductal Papillomas of the Breast	411	53.2	9.5	9.5	12.1	73.0%
	24	Anal Cancer Survivor Rates	536	58.4	8.5	8.5	11.2	92.0%
	25	Can Rhabdomyosarcoma Be Found Early?	296	56.7	9.4	9.4	11.5	91.0%
	26	Radiation Therapy for Colorectal Cancer	1172	48.7	10.2	10.2	12.1	58.0%
	27	Signs and Symptoms for Rhabdomyosarcoma	372	68.2	7.9	7.9	10.5	89.0%
	28	Surgery to Treat Lung Carcinoid Tumors	832	57.5	9.3	9.3	12.5	83.0%
	29	Cancer of Unknown Primary Stages	590	67	7.8	7.8	10.4	100.0%
	30	Targeted Therapy Drugs for Acute Myeloid Leukemia (AML) - ACS Appension	1657	57.9	9.3	9.3	11.6	66.7%
	31	Questions to Ask Your Doctor About Small Intestine Cancer	585	78.6	4.7	4.7	8.0	91.0%
	32	Hormone Therapy for a Cancer of Unknown Primary	353	41.3	11.0	11.0	13.7	73.0%
	33	Treatment for Uterine Sarcoma, by Type and Stage	1314	50.3	10.0	10.0	12.4	67.0%
	34	Protecting People in Clinical Trials	795	58.3	8.5	8.5	11.5	82.0%
	35	Precision or Personalized Medicine	2243	62.5	8.1	8.1	10.7	67.0%
	36	Social Security Disability Insurance for People with Cancer	846	54.1	8.5	8.5	10.9	91.0%
	37	If You Have Laryngeal or Hypopharyngeal Cancer	3133	74.6	6.0	6.0	8.6	87.0%
	38	HPV Signs and Symptoms	576	70.1	7.2	7.2	10.4	73.0%
	39	Cancer Facts for Men	1592	65.2	7.2	7.2	10.0	83.0%
	40	American Cancer Society Recommendations for Prostate Cancer Early Detection	427	54.2	10.1	10.1	12.2	91.0%

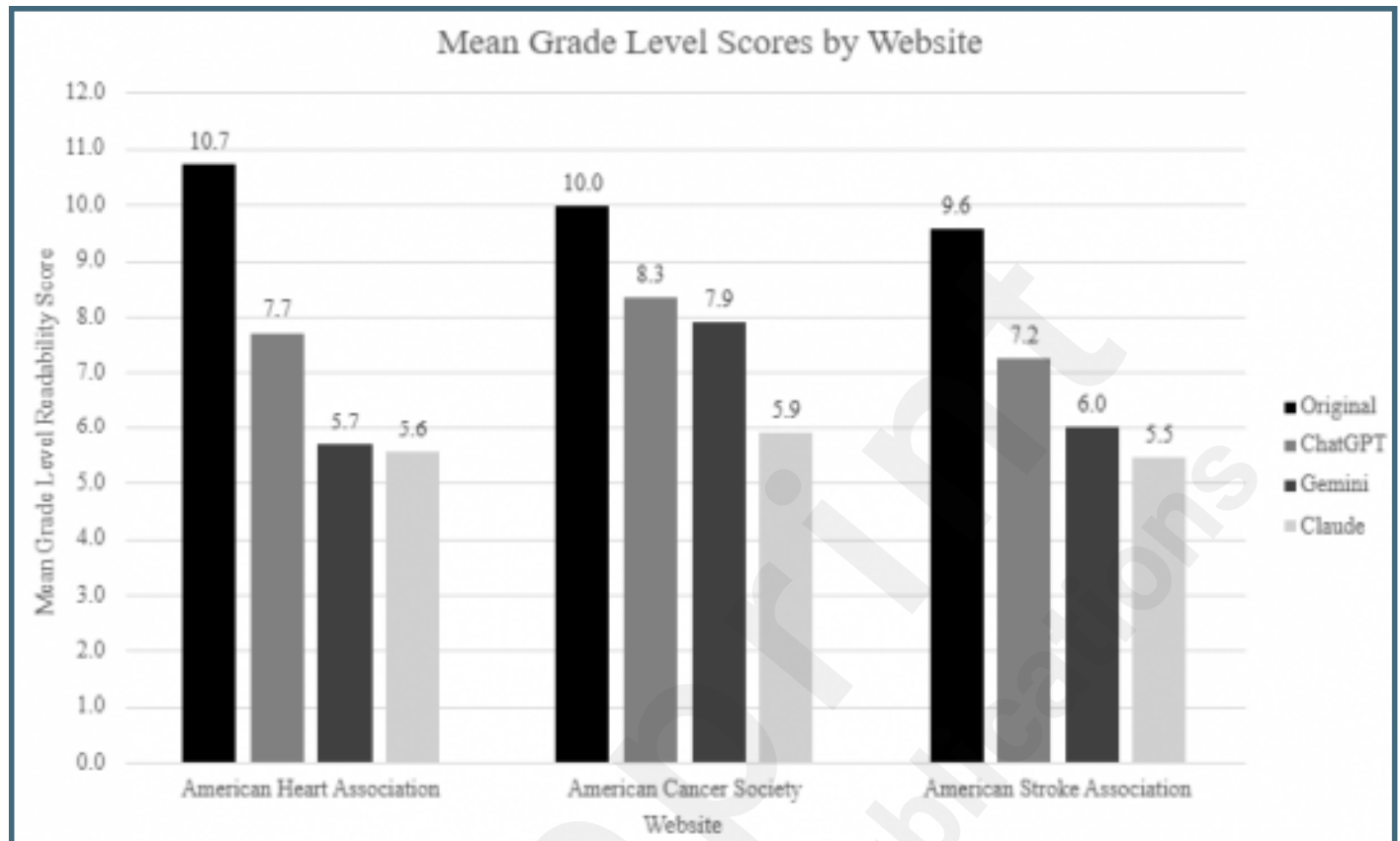
American Stroke Association	41	Hemiparesis	635	48.6	8.6	8.6	10.6	82.0%
	42	Cryptogenic Stroke or Stroke of Unknown Cause	354	47.7	10.0	10.0	12.6	92.0%
	43	15 Things Caregivers Should Know After a Loved One Has Had a Stroke	665	49.3	9.4	9.4	13.1	84.6%
	44	Trouble Swallowing After Stroke (Dysphagia)	480	63.8	6.8	6.8	9.9	83.3%
	45	Choosing the Right Stroke Rehab Facility	929	39.3	10.7	10.7	12.6	87.5%
	46	Effects of Stroke	445	66.4	6.6	6.6	9.5	80.0%
	47	Auditory Overload	157	53.5	7.9	7.9	10.5	83.3%
	48	Brain Stem Stroke	296	57.8	8.9	8.9	11.8	60.0%
	49	Talking to Your Doctor	607	66.6	6.8	6.8	10.3	80.0%
	50	Hearing Disturbances After Stroke	205	53.3	7.9	7.9	10.4	83.3%
	51	Spatial Neglect	323	66.3	7.5	7.5	10.1	58.3%
	52	Bathing Tips for Stroke Survivors	533	82.1	4.3	4.3	7.4	86.7%
	53	Depression and Stroke	337	41.2	9.9	9.9	12.4	66.7%
	54	Aphasia Therapy When Insurance Stops	282	53.1	9.0	9.0	12.2	66.7%
	55	Claw Toe	511	60.9	7.9	7.9	11.0	75.0%
	56	Aspirin and Stroke	529	54	8.9	8.9	11.9	80.0%
	57	Palliative Care	283	36.6	12.1	12.1	14.6	66.7%
	58	Shopping Tips for Stroke Survivors	125	77.1	5.1	5.1	8.8	80.0%
	59	One-Sided Neglect	425	69.1	6.9	6.9	10.5	75.0%
	60	Goal Setting Worksheet	96	94.1	1.8	1.8	6.2	80.0%

FRE Flesch Reading Ease, *FKGL* Flesch-Kincaid Grade Level, *GFI* Gunning-Fog Index, *SMOGI* Simple Measure of Gobbledygook Index, *PEMAT-U* Patient Education Materials Assessment Tools - Understandability

Supplementary Files

Figures

Mean grade level scores for the original version of the patient education material and each large language model version (N = 60).



Multimedia Appendixes

Strobe reporting guidelines checklist for cross-sectional studies.

URL: <http://asset.jmir.pub/assets/005616a4978385679a0a05c13ec273dc.pdf>

TRIPOD-LLM Checklist.

URL: <http://asset.jmir.pub/assets/f00b1a2480ce15cd16ac73e1efd62351.pdf>

