Title

Source Characteristics Influence AI-Enabled Orthopaedic Text Simplification: Recommendations for the Future.

Permalink

https://escholarship.org/uc/item/3vp0k7sx

Journal

JBJS Open Access, 10(1)

Authors

Andalib, Saman
Solomon, Sean
Picton, Bryce
et al.

Publication Date

2025

DOI

10.2106/JBJS.OA.24.00007

JB&JS
OPEN ACCESS

Source Characteristics Influence AI-Enabled Orthopaedic Text Simplification

Recommendations for the Future

Saman Andalib, BS, Sean S. Solomon, BS, Bryce G. Picton, BS, Aidin C. Spina, BS, John A. Scolaro, MD, and Ariana M. Nelson, MD

Investigation performed at the University of California, Irvine, School of Medicine, Irvine, California

Background: This study assesses the effectiveness of large language models (LLMs) in simplifying complex language within orthopaedic patient education materials (PEMs) and identifies predictive factors for successful text transformation.

Methods: We transformed 48 orthopaedic PEMs using GPT-4, GPT-3.5, Claude 2, and Llama 2. The readability, quantified by the Flesch-Kincaid Reading Ease (FKRE) and Flesch-Kincaid Grade Level (FKGL) scores, was measured before and after transformation. Analysis included text characteristics such as syllable count, word length, and sentence length. Statistical and machine learning methods evaluated the correlations and predictive capacity of these features for transformation success.

Results: All LLMs improved FKRE and FKGL scores ($p < 0.01$). GPT-4 showed superior performance, transforming PEMs to a seventh-grade reading level (mean FKGL, 6.72 ± 0.99), with higher FKRE and lower FKGL than other models. GPT-3.5, Claude 2, and Llama 2 significantly shortened sentences and overall text length ($p < 0.01$). Importantly, correlation analysis revealed that transformation success varied substantially with the model used, depending on original text factors such as word length and sentence complexity.

Conclusions: LLMs successfully simplify orthopaedic PEMs, with GPT-4 leading in readability improvement. This study highlights the importance of initial text characteristics in determining the effectiveness of LLM transformations, offering insights for optimizing orthopaedic health literacy initiatives using artificial intelligence (AI).

Clinical Relevance: This study provides critical insights into the ability of LLMs to simplify complex orthopaedic PEMs, enhancing their readability without compromising informational integrity. By identifying predictive factors for successful text transformation, this research supports the application of AI in improving health literacy, potentially leading to better patient comprehension and outcomes in orthopaedic care.

Since the release of ChatGPT by OpenAI, an expanding body of literature intending to assess the implementation of large language models (LLMs) in medicine has emerged. Early iterations of work in this field shed light on the publicly available models' medical knowledge and clinical utility[1-7]. Newer work has transitioned to exploring how LLMs can combat low health literacy[8]. In that exploration, the models have been subjected to quantitative analyses examining their success as a text transformer to restructure complex medical text to a format that is more accessible to patient populations[9-12]. This application of LLM technology has immense potential, given the well-documented discrepancy between the readability of patient-facing documents and the mean reading-comprehension level of their intended populations[13,14]. The success of these transformations suggests a broad applicability of LLMs to transform medical text. The usage of LLMs for this purpose may have the potential to enhance outcomes, as patient understanding of postoperative directives has been shown to facilitate recovery[15-19].

Research with regard to the implementation of LLMs in orthopaedics has been expanding, with a focus on its potential

Disclosure: No external funding was received for this work. The Article Processing Charge for open access publication was funded by University of California, Irvine. The **Disclosure of Potential Conflicts of Interest** forms are provided with the online version of the article (http://links.lww.com/JBJSOA/A729).

as a text transformer to address health literacy[5,20-22]. For patient education materials (PEMs) in the field of spinal surgery, ChatGPT successfully reduced text complexity[21]. Other authors found similar success in transforming online orthopaedic PEMs and have recommended the wide adoption of this technology for this use case[22]. Leveraging this new application of LLMs could be transformative, given the extensive use of PEMs for the orthopaedic surgical population and the relative lack of progress in improving these tools[23-25].

Despite the encouraging results of early analyses of LLM usage in text simplification, further work is needed. The pace of development of new LLMs is remarkably rapid, and previously available models are continually being updated. Academia must adapt quickly to examine these models, as they may present differences in task-specific performance.

Additionally, no studies have examined the textual factors that underlie and predict successful artificial intelligence (AI)-driven text simplification, to our knowledge. An examination of these factors can guide the future clinical implementation of these tools for medical text sources, including PEMs.

Materials and Methods

Data Acquisition

PEMs published online by the American Association of Hip and Knee Surgeons were compiled ($n = 48$). The Flesch-Kincaid Grade Level (FKGL) and Flesch-Kincaid Reading Ease (FKRE) scores, word count, sentence count, mean number of syllables per word, and mean number of syllables per sentence were obtained using the Textstat and NumPy Python packages[26,27]. PEMs were then manually transformed using a standardized prompt in GPT-3.5 (https://chat.openai.com/; OpenAI), GPT-4 (https://chat.openai.com/; OpenAI), Claude 2 (https://claude.ai/; Anthropic), and Llama 2 (https://ai.meta.com/llama/; Meta) between July 27, 2023, and August 1, 2023. The standardized utilized prompt was, "Please rewrite this text to be readable at a fifth-grade level. Do not include information not contained in the original text, and do not exclude information in the original text." The same metrics were then calculated with identical methods for the transformed text.

Descriptive and Inferential Statistics

Data trends were explored for descriptive and inferential statistics using SPSS (Version 29.0.2.0; IBM), GraphPad Prism (Dotmatics), and Matplotlib (Python Software Foundation) plotting libraries. Comparative analysis of quantifiable text qualities was done using single-factor analysis of variance (ANOVA) followed by Games-Howell post hoc analysis.

Transformation Assessment

Transformed PEMs were qualitatively reviewed for obvious extraneous information by 3 authors (S.A., S.S.S., and J.A.S.). For quantitative analyses, latent semantic analysis (LSA) values were calculated using the nltk and sklearn Python packages[28,29]. Preprocessing consisted of lemmatization and the removal of stop-words. The maximum topic occurrence was set as 0.5, and the minimum was set as 1 with 48 total components. The maximum number of features was set to 1,000.

Feature Analysis

Feature analysis was manually performed using random-forest regression, a supervised machine learning method, with n_ estimators, the parameter for the number of decision trees, standardized at 100. This was accomplished with the Pandas, NumPy, and sklearn Python packages[26,28,30]. The input features for the analysis included the original FKGL, original FKRE, original mean syllables per sentence, original mean syllables per word, original number of sentences, and original number of words. The analysis was conducted iteratively for 8 different outputs, representing the post-transformation FKGL and FKRE changes for each of the 4 language models. To ensure a robust assessment of feature importance, we used fivefold cross-validation in each iteration. This process entailed dividing the data into 5 subsets and fitting the model 5 times, each with a different subset reserved as the test set. The reported feature importance values represent the calculated mean across the five folds.

Results

Improving Readability Scores

The original PEMs ($n = 48$) had a mean FKRE (and standard deviation) of 53.40 ± 8.00 corresponding to a tenth-grade reading level (mean FKGL, 10.84 ± 1.49) (see Appendix Supplemental Table 1). On average, GPT-4 increased the FKRE by 23.4 ($p < 0.001$) and reduced the FKGL by 4.08 ($p < 0.001$) (Table I). GPT-3.5 increased the mean FKRE by 16.0 ($p < 0.001$) while reducing the FKGL by 3.15 ($p < 0.001$) (Table I). Claude 2 increased the mean FKRE by 7.4 ($p < 0.01$) and reduced the FKGL by 2.76 ($p < 0.001$). Llama 2 increased the mean FKRE by 10.2 ($p < 0.01$) while reducing the FKGL by 2.11 ($p < 0.001$) (Table I). Comparative analysis (single-factor ANOVA followed by Games-Howell post hoc testing) showed that GPT-4 was best able to improve PEM readability, reaching roughly a seventh-grade reading level on average (6.72 ± 0.99) (see Appendix Supplemental Table 1). There were significant differences in the FKGL change ($p < 0.001$) (Fig. 1-A) and the FKRE change ($p < 0.01$) (Fig. 1-B) between GPT-4 and each of the other LLMs. Comparisons across other text characteristics are included in Figures 1-C through 1-F.

The informational integrity of the PEMs before and after transformation was analyzed using pairwise LSA and was

TABLE I Changes in Readability After LLM Transformation*

Metric	GPT-4	GPT-3.5	Claude 2	Llama 2
FKGL difference	−4.08†	−3.15†	−2.76†	−2.11†
FKRE difference	23.4†	16.0†	7.40‡	10.2‡

*Mean differences in the FKGL and FKRE scores between the original and LLM-transformed PEM. Negative FKGL and positive FKRE differences indicate text simplification. †P < 0.001. ‡P < 0.01.

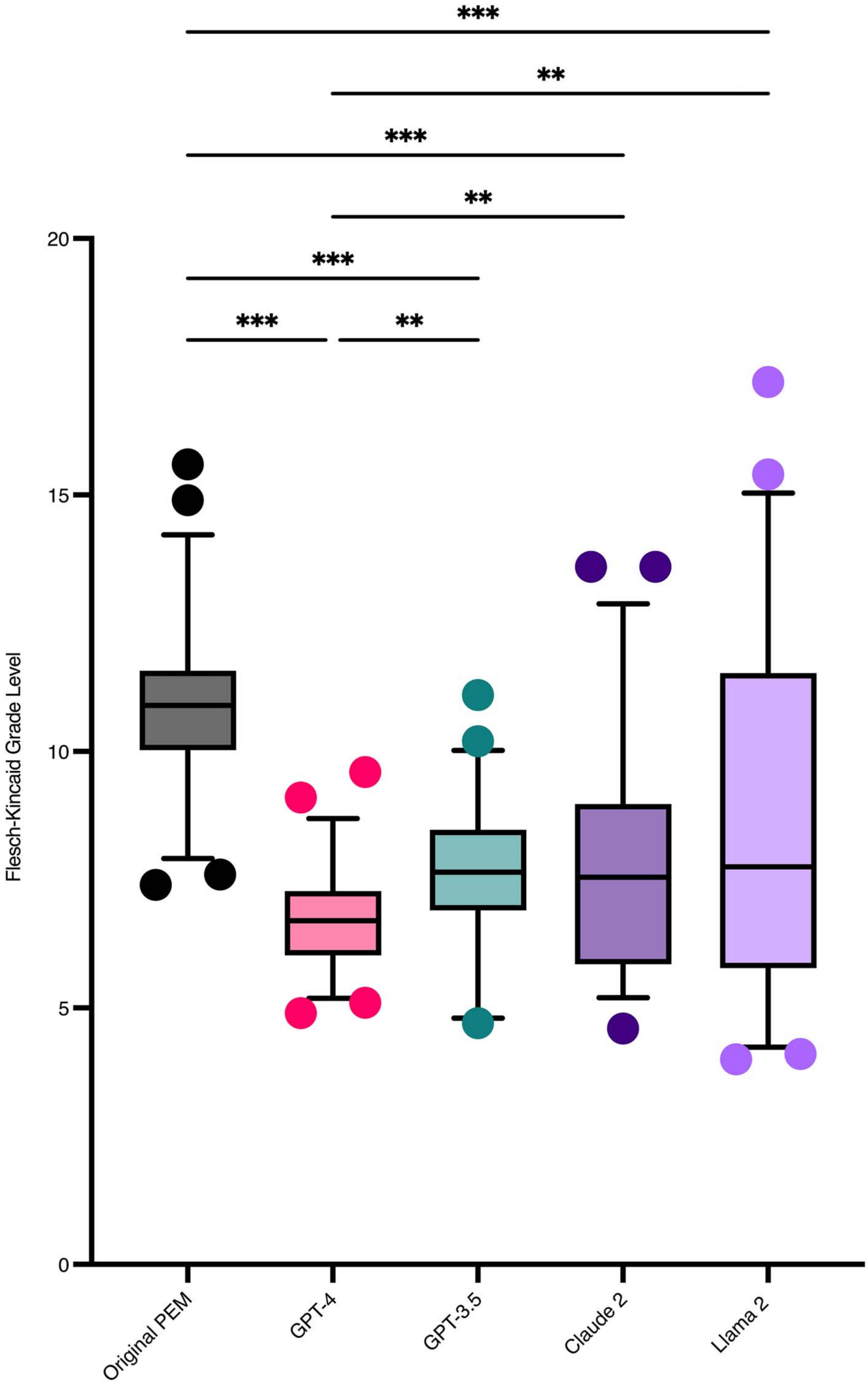

Fig. 1-A

Figs. 1-A through 1-G Comparison of pre-transformation metrics and metrics after transformation of patient education materials by each large language model. *P < 0.05, **p < 0.01, ***p < 0.001. A box indicates the interquartile range (IQR), the line within the box indicates the median, whiskers indicate points within 1.5 times the IQR width of the box, and circles represent outliers. **Fig. 1-A** FKGL score.

reported as a cosine similarity score between 0 and 1. The mean LSA cosine similarity values were 0.945 ± 0.0910 for GPT-4, 0.968 ± 0.0401 for GPT-3.5, 0.961 ± 0.0316 for Claude 2, and 0.908 ± 0.124 for Llama 2 (Fig. 1-G; see also Appendix Supplemental Table 1). Additionally, a qualitative review by 3 authors established that no extraneous information was included in any of the 48 PEMs when transformed by each LLM.

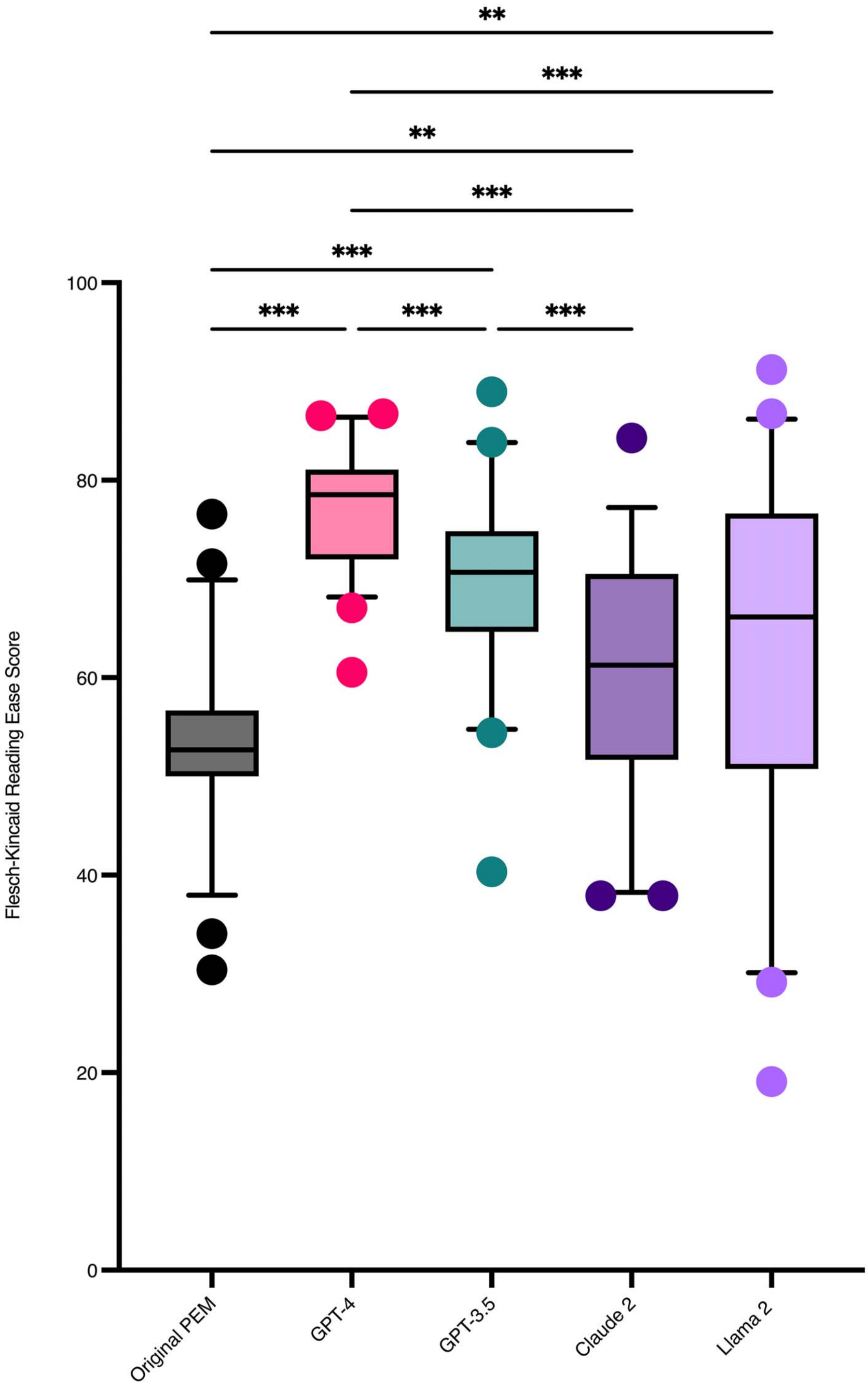

Fig. 1-B

FKRE score.

Correlation Analysis for FKRE Scores

The relationship between original text factors and output FKRE transformation was analyzed using linear regression. A negative correlation was found between the FKGL of the original text and the FKRE of the output text for GPT-4 (−0.38), GPT-3.5 (−0.33), and Claude 2 (−0.35) (Fig. 2-A). Positive correlations were observed between the FKRE score of the original text and the FKRE score of the output text for GPT-4 (+0.42), GPT-3.5 (+0.38), and Claude 2 (+0.40) (Fig. 2-A). The mean syllables per sentence were negatively correlated with the FKRE score of the output text for GPT-4 (−0.29), GPT-3.5 (−0.25), and Claude 2 (−0.28) (Fig. 2-A). The mean syllables per word also

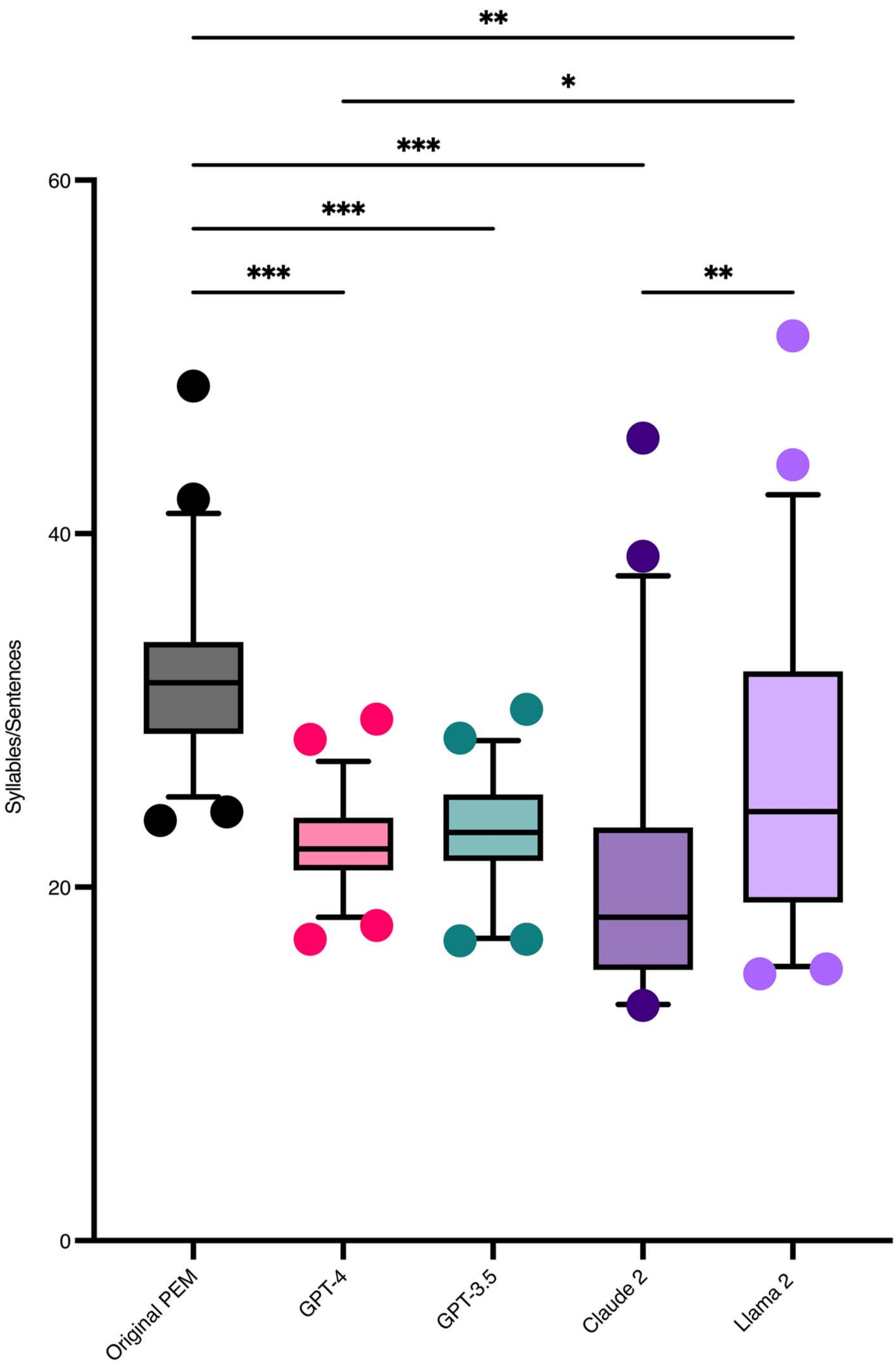

Fig. 1-C
Syllables per sentence.

showed negative correlations with the FKRE score of the output text for GPT-4 (−0.41), GPT-3.5 (−0.35), and Claude 2 (−0.49) (Fig. 2-A). Correlations between the number of sentences and output FKRE score were negative for GPT-4 (−0.18) and GPT-3.5 (−0.34) (Fig. 2-A), but positive for Claude 2 (+0.29) (Fig. 2-A). Similarly, negative correlations were found between the number of words and the output FKRE score for GPT-4 (−0.23) and GPT-3.5 (−0.38) (Fig. 2-A), whereas a positive correlation was observed with Claude 2 (+0.28) (Fig. 2-A). The Llama 2 model did not exhibit any notable correlations between the original text factors and the output FKRE score.

Correlation Analysis for FKGL Scores

The correlation between original text factors and the FKGL of the transformed output was also assessed. The correlation between the FKGL of the original text and the output text

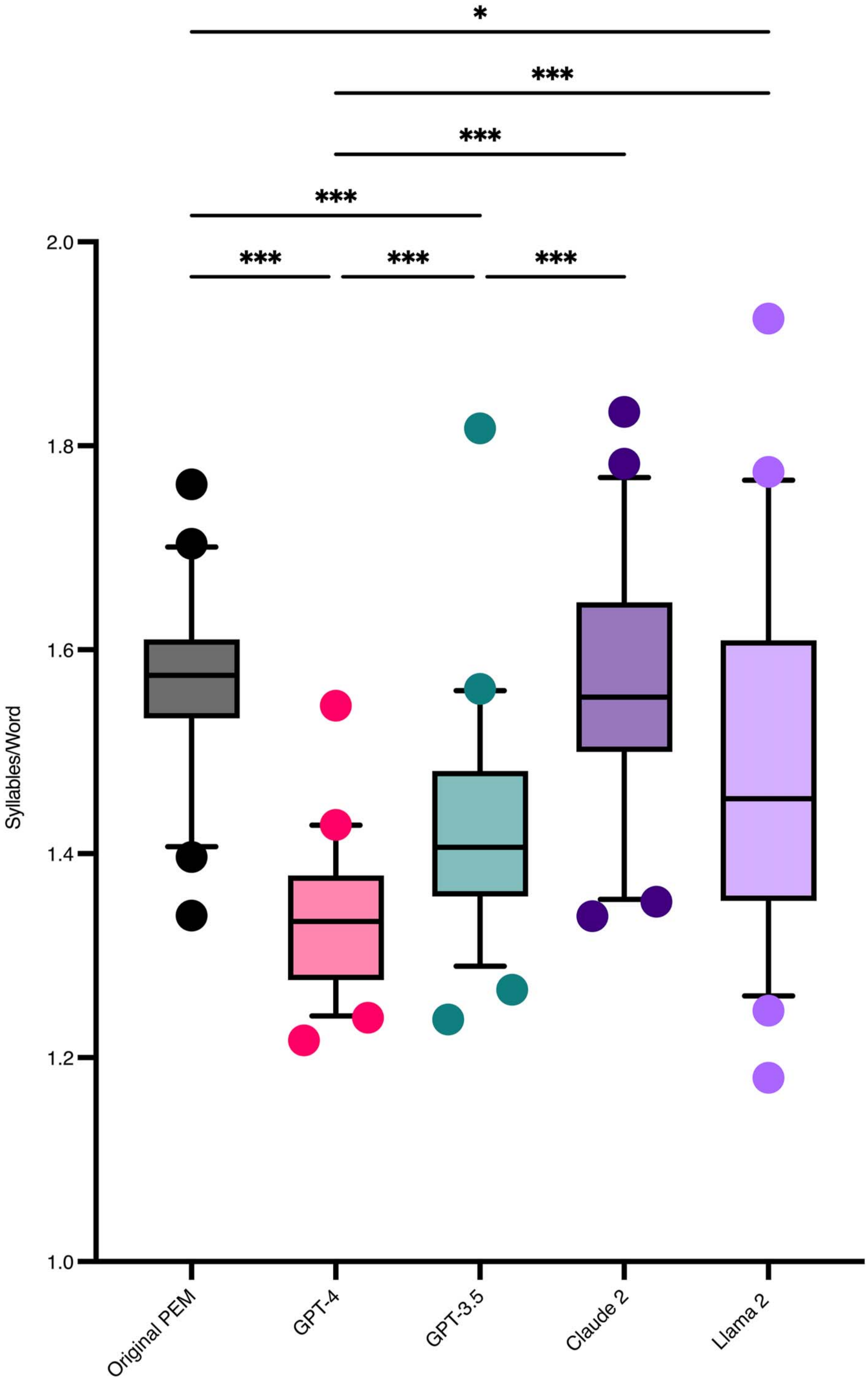

Fig. 1-D
Syllables per word.

yielded a positive correlation for GPT-4 (+0.47), GPT-3.5 (+0.40), and Claude 2 (+0.31) (Fig. 2-B). Additionally, negative correlations were observed between the FKRE score of the original text and the FKGL score of the output text for GPT-4 (−0.43), GPT-3.5 (−0.38), and Claude 2 (−0.31) (Fig. 2-B). For the mean syllables per sentence, there were positive correlations with the FKGL score of the output text for GPT-4 (+0.44), GPT-3.5 (+0.36), and Claude 2 (+0.30) (Fig. 2-B). Likewise, positive correlations were observed between the mean syllables per word and the FKGL score of the output text for GPT-4 (+0.36), GPT-3.5 (+0.31), and Claude 2 (+0.35) (Fig. 2-B). When text-length correlations with the output FKGL were

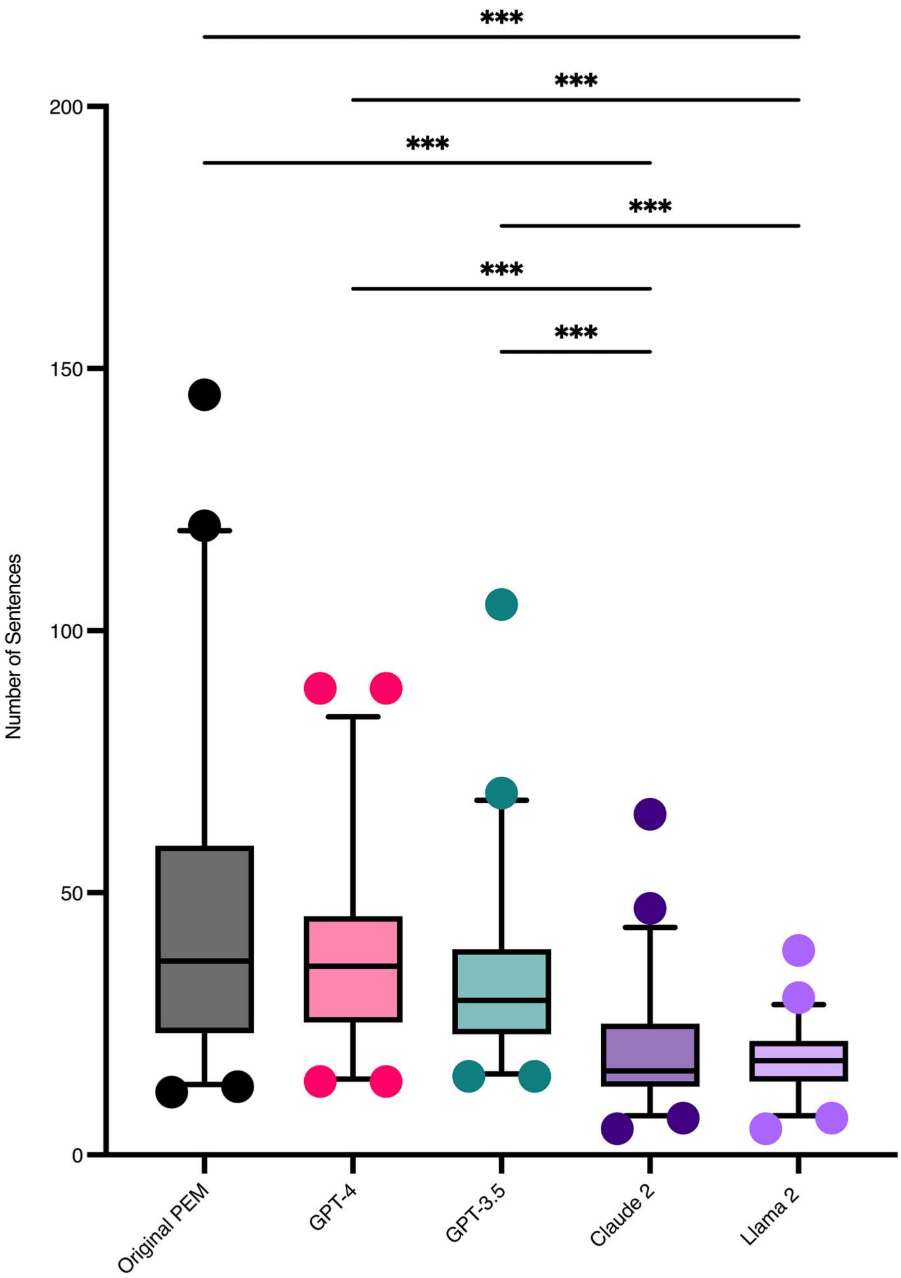

Fig. 1-E
Number of sentences.

quantified, the number of sentences and the output FKRE score had a positive correlation for GPT-3.5 (+0.21) (Fig. 2-B). In contrast, a negative correlation was observed between these variables when using Claude 2 (−0.31) (Fig. 2-B). Correlations between the number of words and the output FKGL scores yielded similar results, as a positive correlation existed between these variables for GPT-3.5 (+0.25) (Fig. 2-B), but a negative correlation existed when using Claude 2 (−0.29) (Fig. 2-B). No significant correlation was found between any of the original text factors and the FKGL score of the output text when using the Llama 2 model.

Feature Analysis

Feature analysis was performed to deduce the relative importance of pre-transformation features in determining a text's post-transformation FKGL and FKRE for each model. The feature importance score means across fivefold cross-validation, based on the differences in FKGL and FKRE resulting from transformation, are shown for each feature in Figures 3-A and 3-B. A factor was defined as significant if it had an importance score of >0.2. The feature importance score means are shown in Table II.

With regard to the post-transformation FKGL, 3 significant feature importance score means were found for GPT-4: the

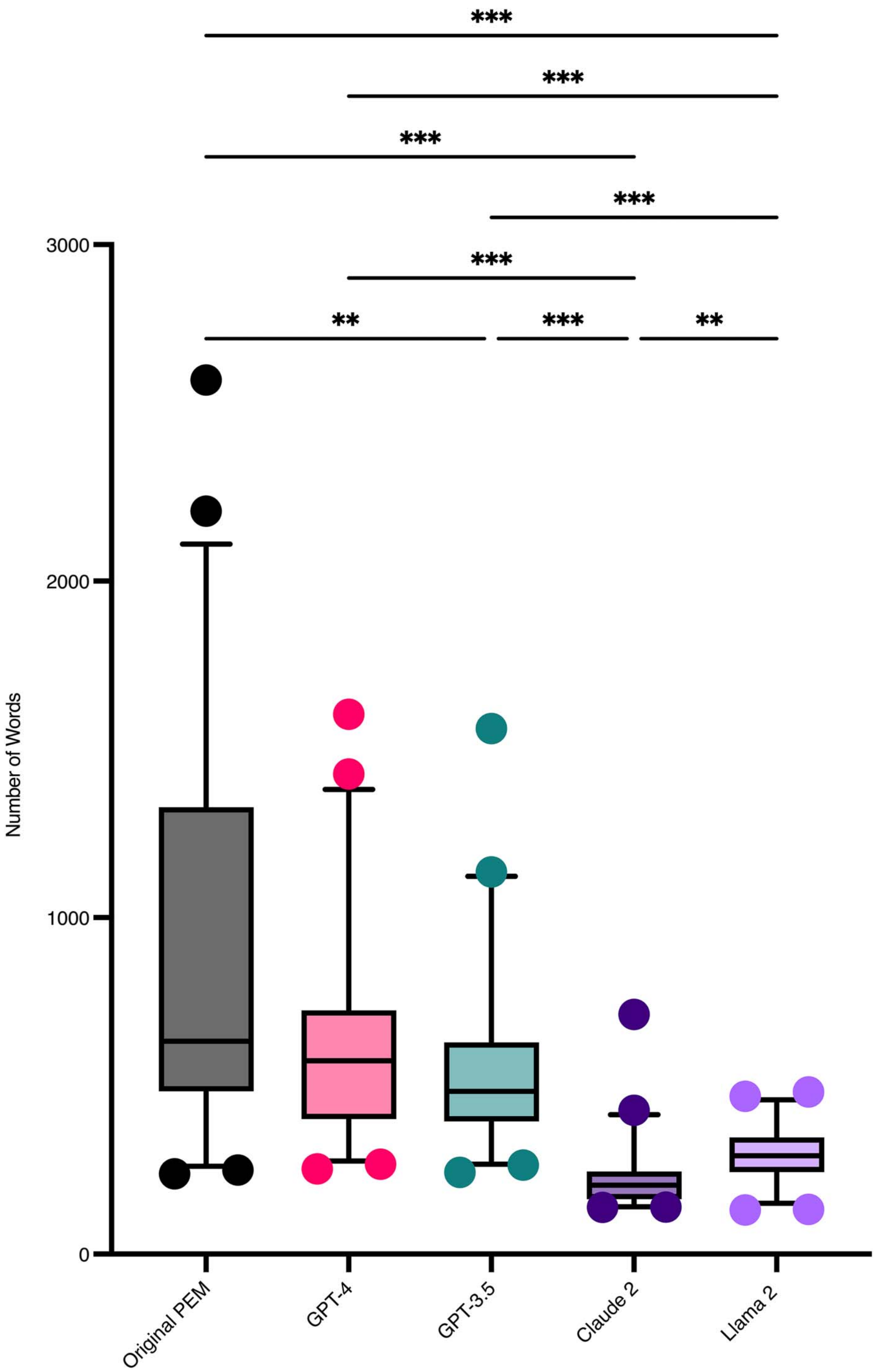

Fig. 1-F
Number of words.

original FKGL (0.2982), original FKRE (0.3190), and original mean syllables per sentence (0.2055) (Fig. 3-A). GPT-3.5 had significant mean scores for only the mean syllables per sentence (0.2765). Claude 2 had significant mean scores for the original FKRE (0.2052) and the original mean syllables per word (0.2226). For Llama 2, mean scores were significant for the original FKGL (0.2137) and the original mean syllables per word (0.2137).

For the post-transformation FKRE, the only significant mean feature importance score for GPT-4 was the original FKRE (0.4566) (Fig. 3-B). GPT-3.5 had significant mean scores for the original FKRE (0.2728) and the original number of sentences (0.2462). Claude 2 had significant mean scores for the original FKRE (0.2784) and original mean syllables per word (0.2141). For Llama 2, the only

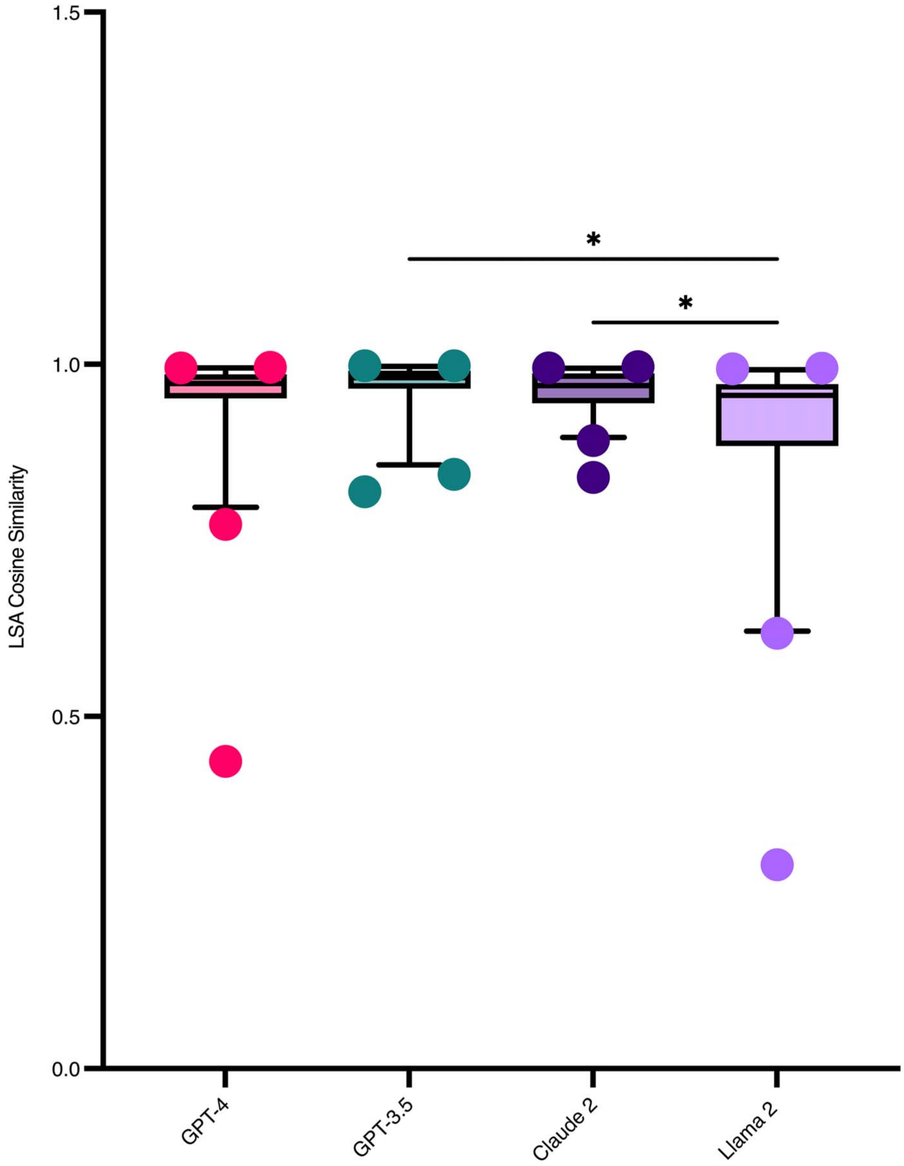

Fig. 1-G
LSA cosine similarity.

significant mean score was the original mean syllables per word (0.3070).

Discussion

LLM Performance

All LLMs were successful in reducing the complexity of orthopaedic PEMs. However, GPT-4 outperformed other models in reducing PEM FKGL scores and increasing PEM FKRE scores (low FKGL and high FKRE indicate simpler text). GPT-4 also maintained the PEM sentence and word length following transformation more consistently than other models. Thus, although other LLMs can achieve relative success, GPT-4 is best able to reduce the complexity of orthopaedic PEMs while maintaining other textual metrics at a consistent level.

Although LLMs can simplify PEMs, post-transformation readability that was at or below a sixth-grade level was not achieved. Other publications have demonstrated the ability of LLMs to transform PEMs to a sixth-grade reading level[21,22]. In contrast to this manuscript, those previous publications did not report the explicit prompts utilized or provide quantitative assessments to ensure post-transformation content integrity. Additionally, those publications did not compare the success of readability transformations across models.

Our prompt's inclusion of the clause "Do not include information not contained in the original text, and do not exclude information in the original text" may explain the observed limitations in transformation success. This clause ensured that PEM information was maintained and external information was not added. LSA and qualitative review were utilized to measure success

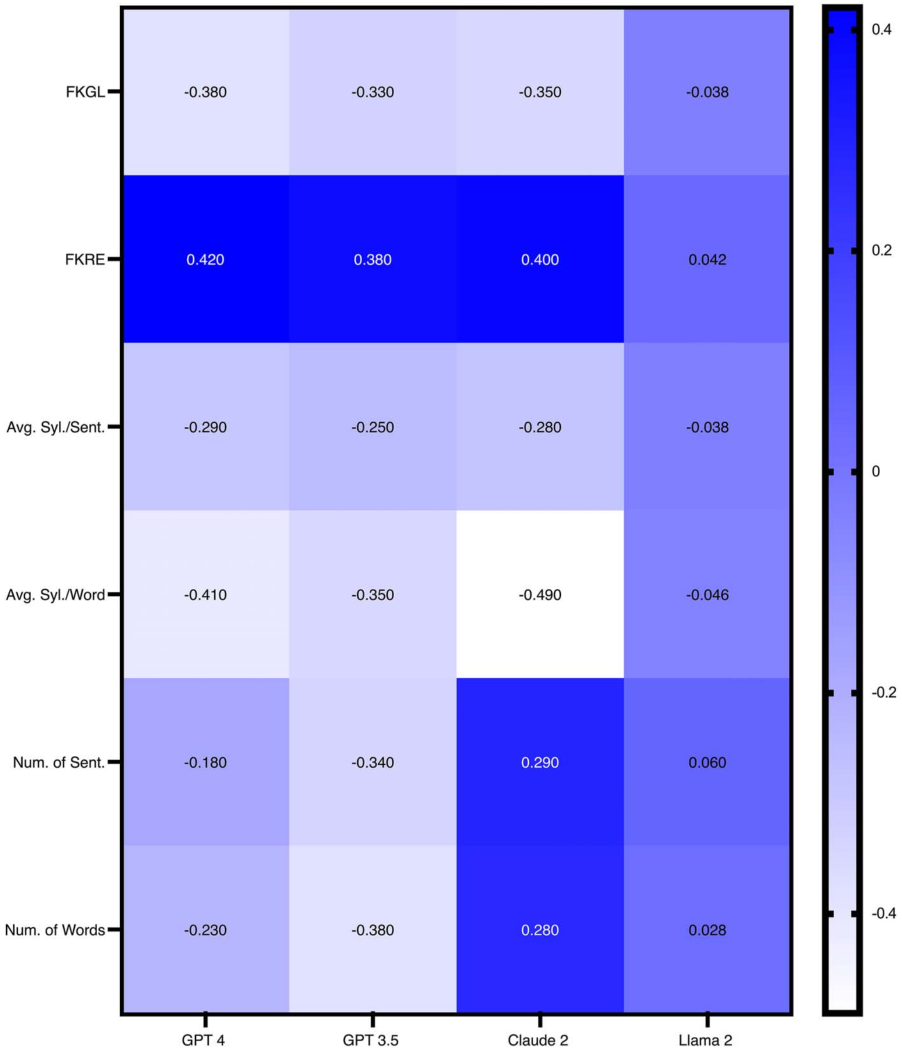

Fig. 2-A

Figs. 2-A and 2-B Correlations of characteristics of the original patient education material with readability of the transformed output. A box indicates the interquartile range (IQR), the line within the box indicates the median, whiskers indicate points within 1.5 times the IQR width of the box, and circles represent outliers. Avg. = average, Syl. = syllable, Sent. = sentence, and Num. = number. **Fig. 2-A** Correlations between original text characteristics and the FKGL of the output generated by each large language model.

in maintaining information integrity[31,32]. LSA and related semantic processing techniques have been previously utilized clinically to assess textual similarity[33,34]. The similarity values of >0.9 for each LLM in our study provide quantitative assurance that core PEM text content was maintained. Thus, the differences in quantitative transformation success may be explained by greater emphasis on content integrity.

Correlations between the initial text characteristics and transformed readability were noted and were variable across the LLMs assessed. Interestingly, text length (as quantified by word and sentence count) negatively correlated with the output FKGL for Claude 2 and positively correlated with it for GPT-4 and GPT-3.5. This trend was consistent between output readability metrics, as text length was positively correlated with output FKRE for Claude 2 and negatively correlated with output FKRE for GPT-4 and GPT-3.5. This result is the first data point to suggest that LLMs perform differently in simplifying PEMs as inherent text qualities change. This is notable, as it suggests that different LLMs may be utilized for medical text simplification based on the original text's writing style and length.

Although correlation analysis reveals the simple association between input and output readability, feature analysis provides insight into the relative weights of input features in determining the output values. The importance of individual linguistic features of PEMs in producing each LLM's readability improvements varied by model. We found that the original FKRE had the greatest feature importance for GPT-4 across both readability measures. The feature with the greatest importance for Llama 2 was the

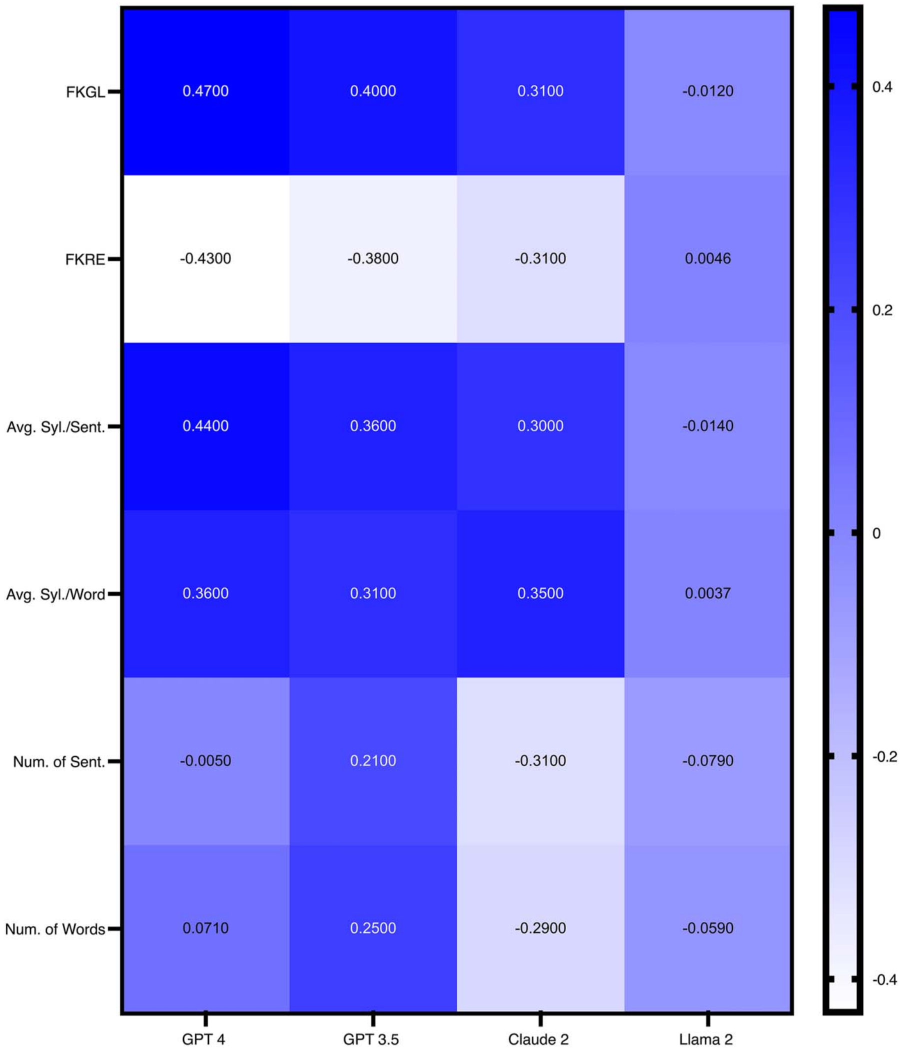

Fig. 2-B
Correlations between original text characteristics and the FKRE of the output generated by each large language model.

original mean syllables per word. No feature was uniformly most important across both readability measures for the other 2 models. These results suggest that the success of LLMs in text simplification is not all-encompassing; instead, it is largely dependent on inherent text qualities. This indicates that physicians may choose to utilize different LLMs for simplifying different text sources.

TABLE II Mean Importance of Original Text Features in Improvement with Transformation by Each Model

	FKGL	FKRE	Mean Syllables per Sentence	Mean Syllables per Word	No. of Sentences	No. of Words
ChatGPT-4 FKGL	0.298	0.319	0.206	0.076	0.054	0.046
ChatGPT-4 FKRE	0.170	0.457	0.072	0.119	0.107	0.073
ChatGPT-3.5 FKGL	0.191	0.171	0.277	0.144	0.116	0.098
ChatGPT-3.5 FKRE	0.111	0.273	0.109	0.139	0.246	0.119
Claude 2 FKGL	0.184	0.205	0.170	0.223	0.118	0.098
Claude 2 FKRE	0.144	0.278	0.133	0.214	0.120	0.109
Llama 2 FKGL	0.214	0.142	0.190	0.214	0.128	0.111
Llama 2 FKRE	0.164	0.147	0.125	0.307	0.129	0.127

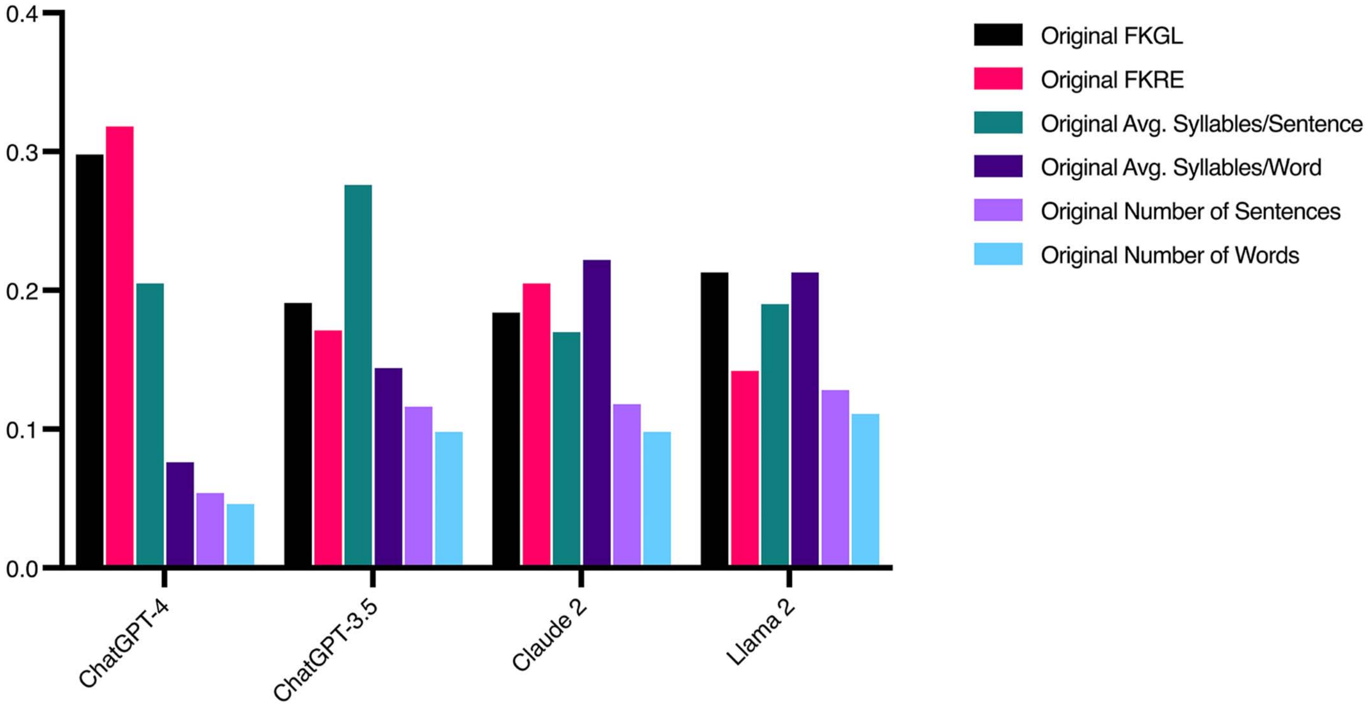

Fig. 3-A

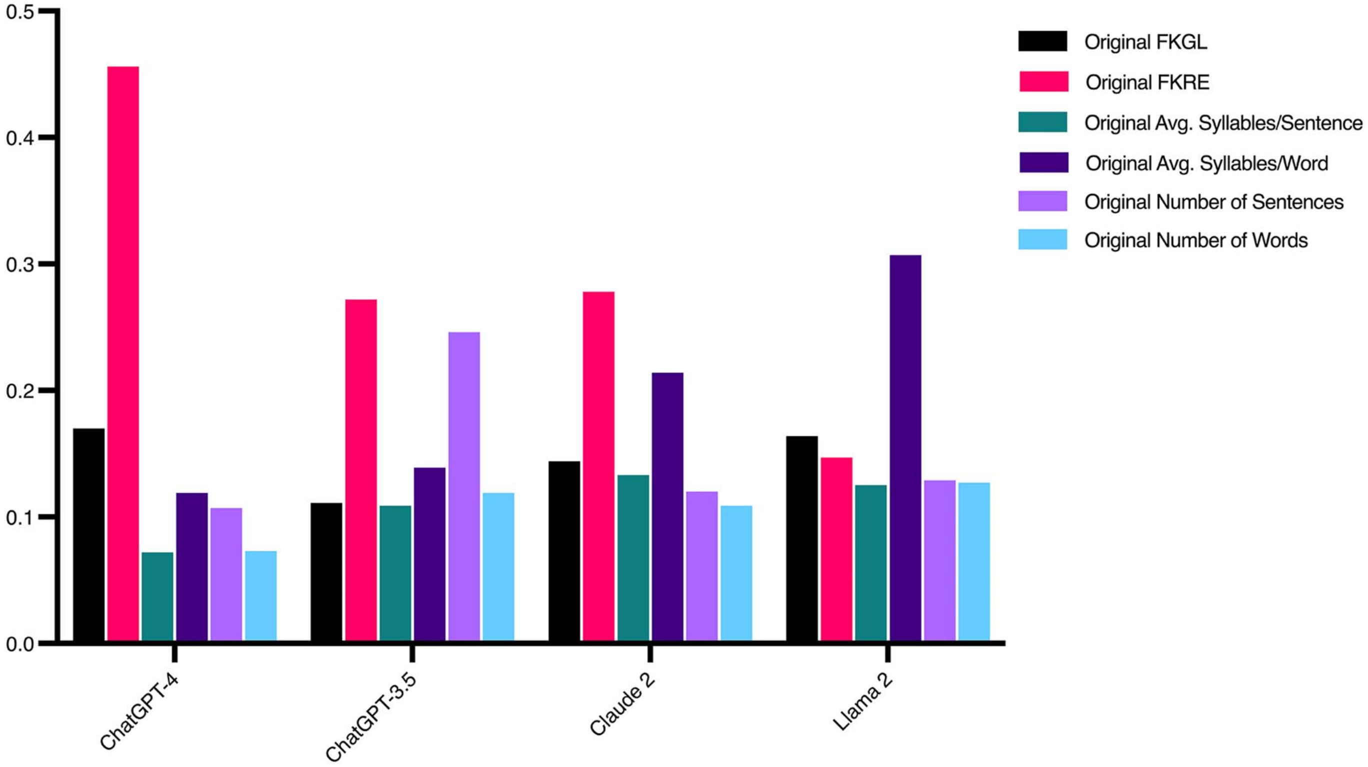

Fig. 3-B

Figs. 3-A and 3-B Importance of each pre-transformation textual feature for successful PEM simplification for LLMs assessed. A box indicates the interquartile range (IQR), the line within the box indicates the median, whiskers indicate points within 1.5 times the IQR width of the box, and circles represent outliers. Avg. = average. **Fig. 3-A** Feature importance scores relative to the difference between the pre-transformation and post-transformation FKGL. **Fig. 3-B** Feature importance scores relative to the difference between the pre-transformation and post-transformation FKRE.

Context and Importance

Regulating the readability of orthopaedic PEMs is essential to ensuring proper patient understanding of preoperative and postoperative directives and has been previously shown to have the potential to improve outcomes[16-19,35-37]. Our results highlight key differences in how modern LLMs transform orthopaedic PEMs to enhance readability. We demonstrated GPT-4's superior performance in maintaining content integrity while sufficiently reducing text complexity. Although all LLMs successfully improved readability metrics without losing core information, only GPT-4 reached an approximately seventh-grade reading level.

Our analysis provides the first evidence that inherent qualities of the original text are differentially associated with each LLM's success. Although text length correlated negatively with the readability of GPT-4 and GPT-3.5 output, positive correlations were observed for Claude 2. Feature analysis enhanced these findings, revealing the relative importance of the original FKRE, word syllables, and sentence length in determining output values.

Together, these results provide a preliminary understanding of how medical texts may need to be prepared to optimize simplification. Materials with higher baseline readability ease and syntactic simplicity may undergo more successful transformation by GPT-4 and GPT-3.5. In contrast, Claude 2 may effectively transform longer, more complex texts. As LLMs advance, these insights will allow providers to select the appropriate model and to tailor materials for simplified patient education. Additionally, our results suggest that regulating certain inherent text qualities can optimize medical text for simplification. Harmonizing publishing protocols using these and future insights holds immense potential for enhancing PEM accessibility, which is one of their existing limitations[38]. Our work constitutes an important early step in unraveling this complex process to pave the way for the widespread implementation of LLMs for this purpose. At present, surgeons should use data-validated prompting techniques as they utilize LLMs to simplify PEMs and other medical text. It is also critical to manually assess PEMs after transformation to ensure fidelity of information and guard against the addition of extraneous prose.

Limitations

This study, although focused, underscores the critical need for further research. We employed FKRE and FKGL readability scores, as they are widely utilized clinical readability metrics that have previously been applied to analysis of PEMs[39,40]. However, future studies could benefit from additional measures for broader complexity analysis[41]. Patient-driven studies should be designed to concurrently validate quantitative results and explore the potential for real-world implementation. The assessment of the success of LLMs in translating PEMs to other languages should also be explored, given their established potential in that context[42].

Although LSA provided valuable data on text similarity, a deeper exploration into the application of that and other semantic analysis techniques is crucial for assessing the reliability and validity of transformations in this burgeoning field[43]. For example, alternative metrics such as explicit semantic analysis, in which textual concepts are labeled rather than assumed as in LSA, may also be utilized to assess transformation success[44]. Although potentially more accurate, however, this approach cannot be automated. Our feature analysis was simplified by assuming feature independence. Future work could explore the interconnected nature of features in real-world data for improved precision. Subsequent studies may also explore restructured prompts or iterative techniques to gauge whether these methods enhance text simplification success and content integrity.

Conclusions

This study evaluates the utility of LLMs to enhance the readability of complex medical texts. We demonstrate that certain LLMs outperform others in simplifying PEMs while maintaining informational integrity. Further, we reveal key correlations between original text characteristics and output reading level, and we conducted feature analysis to identify the metrics that were the most predictive of simplification success. As patients face copious challenges in comprehending health information, this technology holds immense potential to bridge knowledge gaps, with the goal of improving outcomes. Our work unravels the nuances of this process and sets the stage for widespread implementation to benefit patients through enhanced comprehension of texts provided explicitly for their information. Further research will inform the proper utilization of LLMs for medical text simplification in the future.

Appendix

eA Supporting material provided by the authors is posted with the online version of this article as a data supplement at jbjs.org (http://links.lww.com/JBJSOA/A730). ■

Saman Andalib, BS[1]
Sean S. Solomon, BS[1]
Bryce G. Picton, BS[1]
Aidin C. Spina, BS[1]
John A. Scolaro, MD[2]
Ariana M. Nelson, MD[3]

[1]University of California, Irvine, School of Medicine, Irvine, California

[2]Department of Orthopaedic Surgery, University of California, Irvine, Medical Center, Orange, California

[3]Department of Anesthesiology, University of California, Irvine, Medical Center, Orange, California

Email for corresponding author: arianamn@hs.uci.edu

References

1. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, Chartash D. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ. 2023 Feb 8;9:e45312.

2. Hopkins AM, Logan JM, Kichenadasse G, Sorich MJ. Artificial intelligence chatbots will revolutionize how cancer patients access information: ChatGPT represents a paradigm-shift. JNCI Cancer Spectr. 2023 Mar 1;7(2): pkad010.

3. Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a radiology board-style examination: insights into current strengths and limitations. Radiology. 2023 Jun;307(5):e230582.

4. Scherr R, Halaseh FF, Spina A, Andalib S, Rivera R. ChatGPT interactive medical simulations for early clinical education: case study. JMIR Med Educ. 2023 Nov 10;9:e49877.

5. Lum ZC. Can artificial intelligence pass the American Board of Orthopaedic Surgery examination? Orthopaedic residents versus ChatGPT. Clin Orthop Relat Res. 2023 Aug 1;481(8):1623-30.

6. Tong WJ, Wu SH, Cheng MQ, Huang H, Liang JY, Li CQ, Guo HL, He DN, Liu YH, Xiao H, Hu HT, Ruan SM, Li MD, Lu MD, Wang W. Integration of artificial intelligence decision aids to reduce workload and enhance efficiency in thyroid nodule management. JAMA Netw Open. 2023 May 1;6(5):e2313674.

7. Patil NS, Huang R, van der Pol CB, Larocque N. Using AI chatbots as a radiologic decision-making tool for liver imaging: do ChatGPT and Bard communicate information consistent with the American College of Radiology Appropriateness Criteria? J Am Coll Radiol. 2023 Jul 28.

8. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel). 2023 Mar 19;11(6):887.

9. Davis R, Eppler M, Ayo-Ajibola O, Loh-Doyle JC, Nabhani J, Samplaski M, Gill I, Cacciamani GE. Evaluating the effectiveness of artificial intelligence-powered large language models application in disseminating appropriate and readable health information in urology. J Urol. 2023 Oct;210(4):688-94.

10. Spina A, Andalib S, Flores D, Vermani R, Halaseh FF, Nelson AM. Evaluation of generative language models in personalizing medical information: instrument validation study. JMIR AI. 2024 Aug 13;3:e54371.

11. Haver HL, Lin CT, Sirajuddin A, Yi PH, Jeudy J. Use of ChatGPT, GPT-4, and Bard to improve readability of ChatGPT's answers to common questions about lung cancer and lung cancer screening. AJR Am J Roentgenol. 2023 Nov;221(5):701-4.

12. Eid K, Eid A, Wang D, Raiker RS, Chen S, Nguyen J. Optimizing ophthalmology patient education via ChatBot-generated materials: readability analysis of AI-generated patient education materials and the American Society of Ophthalmic Plastic and Reconstructive Surgery patient brochures. Ophthalmic Plast Reconstr Surg. 2024 Mar-Apr 01;40(2):212-6.

13. Imoisili OE, Levinsohn E, Pan C, Howell BA, Streiter S, Rosenbaum JR. Discrepancy between patient health literacy levels and readability of patient education materials from an electronic health record. Health Lit Res Pract. 2017 Nov 9;1(4): e203-7.

14. Institute of Medicine (US) Committee on Health Literacy. Nielsen-Bohlman L, Panzer AM, Kindig DA, editors. Health Literacy: A Prescription to End Confusion. Washington (DC): National Academies Press; 2004.

15. Atchison KA, Black EE, Leathers R, Belin TR, Abrego M, Gironda MW, Wong D, Shetty V, DerMartirosian C. A qualitative report of patient problems and postoperative instructions. J Oral Maxillofac Surg. 2005 Apr;63(4):449-56.

16. Goldchmit SM, de Queiroz MC, Dos Anjos Rabelo ND, Junior WR, Polesello GC. Patient education in orthopedics: the role of information design and user experience. Curr Rev Musculoskelet Med. 2021 Feb;14(1):9-15.

17. Sunjaya AP, Bao L, Martin A, DiTanna GL, Jenkins CR. Systematic review of effectiveness and quality assessment of patient education materials and decision aids for breathlessness. BMC Pulm Med. 2022 Jun 20;22(1):237.

18. Thomas ND, Mahler R, Rohde M, Segovia N, Shea KG. Evaluating the readability and quality of online patient education materials for pediatric ACL tears. J Pediatr Orthop. 2023 Oct 1;43(9):549-54.

19. Dykes PC, Carroll DL, Hurley A, Lipsitz S, Benoit A, Chang F, Meltzer S, Tsurikova R, Zuyov L, Middleton B. Fall prevention in acute care hospitals: a randomized trial. JAMA. 2010 Nov 3;304(17):1912-8.

20. Mika AP, Martin JR, Engstrom SM, Polkowski GG, Wilson JM. Assessing ChatGPT responses to common patient questions regarding total hip arthroplasty. J Bone Joint Surg Am. 2023 Oct 4;105(19):1519-26.

21. Baumann J, Marshall S, Groneck A, Hanish SJ, Choma T, DeFroda S. Readability of spine-related patient education materials: a standard method for improvement. Eur Spine J. 2023 Sep;32(9):3039-46.

22. Kirchner GJKR, Kim RY, Weddle JB, Bible JE. Can artificial intelligence improve the readability of patient education materials? Clin Orthop Relat Res. 2023 Nov 1; 481(11):2260-7.

23. Doinn TO, Broderick JM, Abdelhalim MM, Quinlan JF. Readability of patient educational materials in hip and knee arthroplasty: has a decade made a difference? J Arthroplasty. 2020 Nov;35(11):3076-83.

24. Ó Doinn T, Broderick JM, Abdelhalim MM, Quinlan JF; T OD. Readability of patient educational materials in pediatric orthopaedics. J Bone Joint Surg Am. 2021 Jun 16;103(12):e47.

25. Johansson K, Salanterä S, Katajisto J, Leino-Kilpi H. Written orthopedic patient education materials from the point of view of empowerment by education. Patient Educ Couns. 2004 Feb;52(2):175-81.

26. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ, Kern R, Picus M, Hoyer S, van Kerkwijk MH, Brett M, Haldane A, Del Río JF, Wiebe M, Peterson P, Gérard-Marchant P, Sheppard K, Reddy T, Weckesser W, Abbasi H, Gohlke C, Oliphant TE. Array programming with NumPy. Nature. 2020 Sep;585(7825):357-62.

27. 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. 1975. Accessed 2024 Jun 6. https://stars.library.ucf.edu/istlibrary/56/

28. Pedregosa FVG, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011 Nov 1;12(85):2825-30.

29. Bird SKE, Loper E. Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. O'Reilly Media; 2009.

30. McKinney W. Data structures for statistical computing in Python. In: Proceedings of the Ninth Python in Science Conference; 2010. p. 56-61.

31. Vrana SR, Vrana DT, Penner LA, Eggly S, Slatcher RB, Hagiwara N. Latent semantic analysis: a new measure of patient-physician communication. Soc Sci Med. 2018 Feb;198:22-6.

32. Deerwester SDS, Furnas GW, Landauer TK, Harshman R. Indexing by latent semantic analysis. J Am Soc Inf Sci. 1990;41(6):391-407.

33. García AM, Escobar-Grisales D, Vásquez Correa JC, Bocanegra Y, Moreno L, Carmona J, Orozco-Arroyave JR. Detecting Parkinson's disease and its cognitive phenotypes via automated semantic analyses of action stories. NPJ Parkinsons Dis. 2022 Nov 25;8(1):163.

34. Leleu TD, Jacobson IG, LeardMann CA, Smith B, Foltz PW, Amoroso PJ, Derr MA, Ryan MA, Smith TC; Millennium Cohort Study Team. Application of latent semantic analysis for open-ended responses in a large, epidemiologic study. BMC Med Res Methodol. 2011 Oct 5;11:136.

35. Badarudeen S, Sabharwal S. Readability of patient education materials from the American Academy of Orthopaedic Surgeons and Pediatric Orthopaedic Society of North America web sites. J Bone Joint Surg Am. 2008 Jan;90(1):199-204.

36. Beall MS 3rd, Golladay GJ, Greenfield ML, Hensinger RN, Biermann JS. Use of the Internet by pediatric orthopaedic outpatients. J Pediatr Orthop. 2002 Mar-Apr; 22(2):261-4.

37. Krempec J, Hall J, Biermann JS. Internet use by patients in orthopaedic surgery. Iowa Orthop J. 2003;23:80-2.

38. Rooney MK, Santiago G, Perni S, Horowitz DP, McCall AR, Einstein AJ, Jagsi R, Golden DW. Readability of patient education materials from high-impact medical journals: a 20-year analysis. J Patient Exp. 2021 Mar 3;8:2374373521998847.

39. O'Sullivan L, Sukumar P, Crowley R, McAuliffe E, Doran P. Readability and understandability of clinical research patient information leaflets and consent forms in Ireland and the UK: a retrospective quantitative analysis. BMJ Open. 2020 Sep 3; 10(9):e037994.

40. Eltorai AE, Ghanian S, Adams CA Jr, Born CT, Daniels AH. Readability of patient education materials on the American Association for Surgery of Trauma website. Arch Trauma Res. 2014 Apr 30;3(2):e18161.

41. Jindal P, MacDermid JC. Assessing reading levels of health information: uses and limitations of Flesch formula. Educ Health (Abingdon). 2017 Jan-Apr;30(1):84-8.

42. Soto-Chávez MJ, Bustos MM, Fernández-Ávila DG, Muñoz OM. Evaluation of information provided to patients by ChatGPT about chronic diseases in Spanish language. Digit Health. 2024 Jan 2;10:20552076231224603.

43. Suleman R, Korkontzelos I. Extending latent semantic analysis to manage its syntactic blindness. Expert Systems with Applications. 2021;165:114130.

44. Woods DL, Wyma JM, Herron TJ, Yund EW. Computerized analysis of verbal fluency: normative data and the effects of repeated testing, simulated malingering, and traumatic brain injury. PLoS One. 2016 Dec 9;11(12):e0166439.