Evaluation of Commonly Utilized Artificial Intelligence Large Language Models in Optimizing Readability, Accuracy, and Patient Comprehension of Orthopaedic Oncology Patient Educational Materials

Patrick Nian, Christopher Williams, Joydeep Baidya, Isabella G Marsh, Ithika S Senthilnathan, Aditya V Maheshwari

INTRODUCTION: Online patient educational materials (PEMs) help patients and families better understand diagnosis, treatment options, and expected outcomes, but have poor readability, limiting their intended purposes. Orthopaedic oncology related topics are particularly complex and require overlap across multiple medical disciplines. Orthopaedic oncology PEMs have been shown to read > the ninth grade reading level, exceeding the recommended sixth-grade reading level set forth by the American Medical Association. Recent literature has pointed to the limited efficacy of ChatGPT in simplifying the readability of orthopaedic PEMs, but other common large language models (LLMs) have not yet been evaluated for this use. Understanding the most efficacious LLM that does not compromise accuracy or patient comprehension can guide partnerships between healthcare organizations and artificial intelligence companies. Therefore, the aims of the study were (1) to assess the initial readability of online English language PEMs related to orthopaedic oncology and (2) to compare five commonly utilized LLMs in their efficacy in improving readability while maintaining accuracy and patient comprehension.

METHODS: Seventy-two PEMs were collected from academic and professional associations (AAOS [Orthoinfo.org], MSTS, and the top five ranked orthopaedic hospitals). Content directed towards education of health professionals and news articles were excluded. Readability metrics included two grade-level metrics (Flesch-Kincaid Grade Level [FKGL] and Gunning Fog Index [GFI], with higher scores indicating more difficult readability), and the Flesch Reading Ease (FRE, scored 1 to 100, with higher score indicating greater readability). The written content of each PEM was inputted into five LLMs, including ChatGPT-4o, Google Gemini, DeepSeek AI, Microsoft Copilot, and Meta AI, which were prompted to “rewrite this document to a sixth-grade reading level”. The readability of each output was assessed. Two independent graders evaluated the artificial intelligence content for its retention of patient comprehension and accuracy, measured by the F1 score, a composite score of precision and recall through a categorization of converted text as a true positive, false positive, false negative, and true negative (Figure 1). ANOVA tests followed by pairwise comparisons compared each LLM to baseline and other LLMs. Secondary analysis of the 28 Orthoinfo PEMs was performed to assess the readability, accuracy, and patient comprehension when the LLMs were prompted to rewrite to the fifth-, fourth-, and third- grade reading level.

RESULTS: The baseline readability of the 72 PEMs was between the 8th and 9th grade reading level (FKGL: 8.7 ± 1.5), and slightly higher when measured by GFI (GFI: 10.5 ± 1.9). Baseline FRE was 53.9 ± 8.2. Only one PEM (1.0%) from Cedars-Sinai related to giant cell tumor had a baseline readability at or below a sixth grade FKGL. Prompting of the five LLMs resulted in significant improvements in all readability metrics from baseline ($P < 0.001$). ChatGPT-4o, DeepSeek AI, and Google Gemini conversion resulted in the most readable PEMs, which were significantly more readable than MetaAI and Microsoft Copilot. (Figure 2) Google Gemini had the highest F1 score of 0.986 (range: 0.765 to 0.986) and patient comprehension of 100%. (Figure 3) Sub-analyses showed that compared to the sixth-grade prompt, significant improvements in readability were typically achieved when prompted to the third-grade. (Figure 4) Accuracy and patient comprehension was substantially compromised when prompted to lower grade levels for Meta AI. (Figure 5)

DISCUSSION AND CONCLUSION: ChatGPT-4o, Google Gemini, DeepSeek AI were moderately effective in achieving recommended reading levels maintaining high accuracy and patient comprehension. Microsoft Copilot and MetaAI had inferior efficacy, which may reflect the public data that these models are trained on. Commonly utilized LLMs are novel tools to rapidly improve the readability for patient use. These findings may also guide partnership decisions between healthcare organizations and artificial intelligence companies and help consumers choose the right AI LLMs.

Figure 1. Evaluation of F1 Accuracy Score of AI-converted Content

	AI-Converted Content	
Original Content	**True Positive (TP):** AI-converted content is correct and exists in the original document	**False Negative (FN):** Original content contains information that LLM failed to generate
	False Positive (FP): AI-converted content is correct but introduces new content that was not present in the original document	**True Negative (TN):** AI-converted content is incorrect, and the original document contains the correct version
Precision = TP/(TP+FP) Recall = TP/(TP+FN) **F1 score** = 2*(precision*recall)/(precision + recall)