Can Large Language Models Meet Patient Literacy Needs? Readability in Total Hip Arthroplasty Queries.

Paul S Soliman[1], Anoop Chandrashekar, John R Martin
[1]Orthopedics Department

INTRODUCTION: Patients utilize a variety of information sources to better equip themselves in medical decision making. With the advent of Large Language Models (LLMs), there has been increasing reliance on these models as a primary method for obtaining health information. Concurrently, the NIH has mandated that patient-facing resources be written at a 6th grade reading level. Given that many patients will likely turn to LLM models when gathering health information, this paper seeks to evaluate the readability and 'prompt-ability' of popular LLM models, including OpenAI's ChatGPT 4o, Google's Gemini, Meta's Llama, and Deepseek, in response to frequently asked Total Hip Arthroplasty (THA) queries.

METHODS: This study employs a cross-sectional approach to evaluate the readability and prompt-ability of responses from four widely used artificial intelligence chatbots to ten validated queries regarding THA. These queries were derived from prior literature which aggregated commonly asked questions provided by reputable healthcare institutions. Each query was submitted to the chatbot, both with and without prompting, to tailor responses to a 6th grade reading level. This process was replicated three times independently for each chatbot, accruing a total of 60 responses (30 with prompting, 30 without) per model. A novel prompt-ability metric was devised, quantifying the models' ability to adjust linguistic complexity upon request. The performance across chatbots was then compared.

RESULTS: Upon analysis, all chatbot models exhibited an enhancement in readability when prompted, with Google's Gemini demonstrating the greatest improvement of 4.79 Flesch Kincaid Grade levels (12.25 to 7.46). However, none of the models were able to achieve a 6th grade reading level. Inter-model comparisons revealed statistically significant differences in the average readability of these models both when prompted and unprompted. Differences in the 'prompt-ability' of models was also demonstrated, with Google's Gemini exhibiting the greatest prompt-ability and Meta's Llama exhibiting the least.

DISCUSSION AND CONCLUSION: In assessing the readability of LLM chatbot responses, this study found that while all models exhibit enhanced readability when prompted, the extent of improvement does not align with the readability standards set by the NIH. Furthermore, this research uncovered discernible disparities among the models in their ability to diminish linguistic complexity when prompted.