

AI's Might vs Human Prowess in Crafting Plain Language Summaries

Sonali Satam¹, Shweta Pitre², Priya Ganpathy², Madhavi Patil¹, Vatsal Shah³

¹SIRO Clinpharm Pvt. Ltd, India; ²SIRO Clinpharm UK Limited, London, UK; ³SIRO Clinpharm USA LLC, Princeton, NJ, USA.

INTRODUCTION

- Plain language summaries (PLS) of publications foster health literacy and empower patients and other non-specialist audiences by providing access to simplified and digestible, but credible scientific information.^{1,2}
- Artificial intelligence (AI) is emerging as a formidable tool in scientific writing,^{3,4} however, its ability to meet the rigorous standards of biomedical writing and grasp the nuances of health literacy for PLS development require a detailed assessment.

OBJECTIVE

- This study was designed to evaluate if ChatGPT generated PLS (ChatGPT-PLS) can match medical writer (MW) generated PLS (Human-PLS) in overall quality and scientific rigor.

METHODS

Data collection

- 15 Human PLS:** Full-length published standalone oncology PLS developed by skilled medical writers (MW) were selected.
- 15 ChatGPT-PLS:** PLS of the same oncology primary studies (selected above) were generated using ChatGPT AI tool (v3.5).
- All 30 PLS** were blinded by an independent resource; MW (n=2) experienced in PLS development and lay reviewers LR (n=2) evaluated the PLS.

Assessment

- The human-PLS and ChatGPT-PLS were analyzed using a four-point Likert scale for
 - Accuracy
 - Completeness
 - Clarity and readability
 - Conciseness
 - Coherence
 - Overall quality
- Readability was assessed using the Flesch-Kincaid scores and Gunning Fog Index.
- MW also conducted a content review comparing both PLS types to their primary publications to evaluate the scientific rigor.

Statistical analysis

- MW and LR rated each outcome on a 4-point Likert scale, ranging from Excellent (4) to Poor (1).
- The mean difference for each outcome between Human-PLS and ChatGPT-PLS, as well as the mean difference between the readability scores, were assessed using the Mann-Whitney U test.

Human-PLS were better than ChatGPT-PLS in accuracy, completeness, and conciseness, underscoring the need for human intervention in scientific content creation.

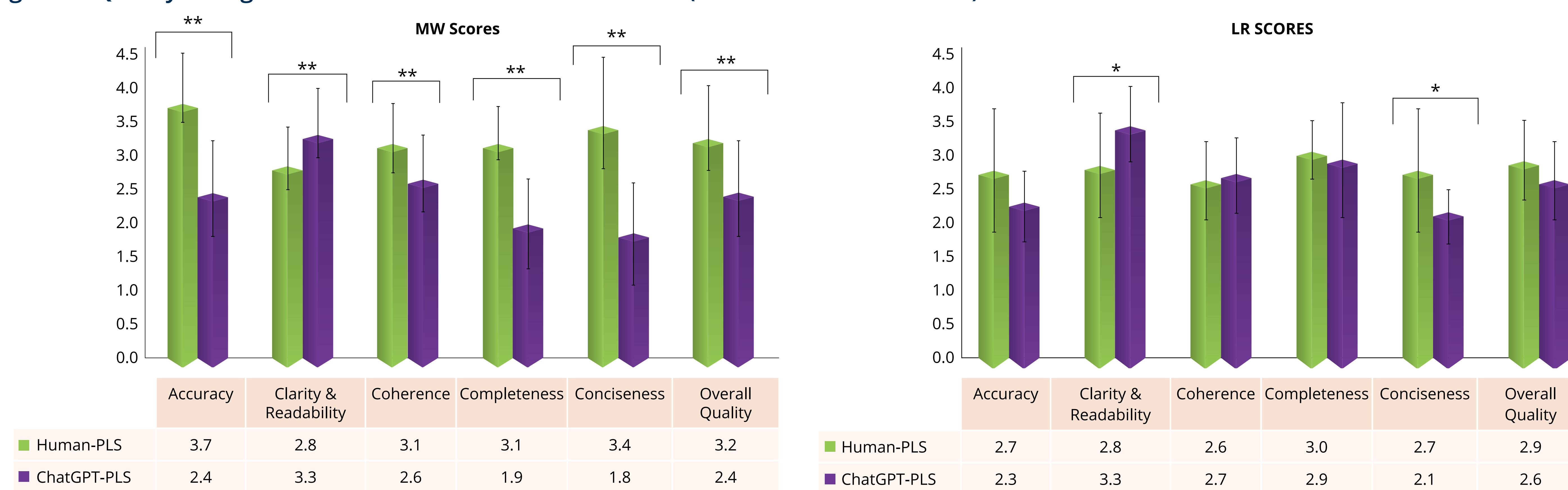
ChatGPT enhances readability of PLS, suggesting the potential of AI in aiding PLS development with accessible medical information. However, human review is essential to ensure ChatGPT doesn't compromise the accuracy/precision of PLS.

The might of AI to enhance readability along with human prowess to uphold scientific rigor should work in synergy for efficient PLS development.

RESULTS

- Human-PLS scored significantly higher ($p < 0.01$) on accuracy, coherence, completeness, conciseness, and overall quality versus ChatGPT-PLS.
- ChatGPT-PLS had significantly higher ($p < 0.01$) clarity and readability versus Human-PLS.
- Both MW and LR observations were generally consistent across the study outcomes.

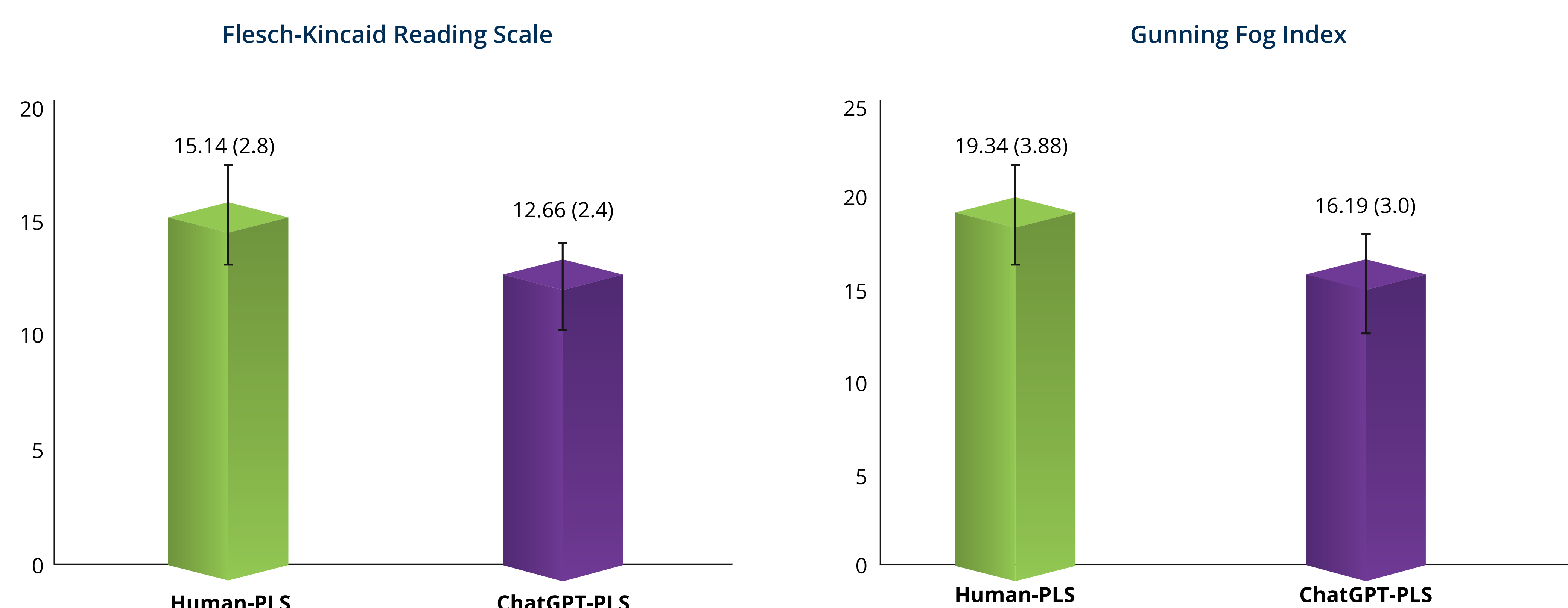
Figure 1: Quality ratings for Human-PLS versus ChatGPT-PLS (MW- rated and LR-rated)



*pvalue<0.05; **pvalue<0.01; PLS evaluation outcomes based on above parameters were scored using 4-point Likert scales (Excellent=4; Good=3; Fair=2; Poor=1). LR, lay reviewers; MW, medical writers; PLS, plain language summaries

- ChatGPT-PLS were easier to read compared to Human-PLS (Flesch-Kincaid Reading Scale: $p=0.020$; Gunning Fog Index: $p=0.015$).
- However, both ChatGPT-PLS and Human-PLS were of moderate complexity, with a reading level suitable for college seniors and graduates.

Figure 2: Readability scores for Human-PLS versus ChatGPT-PLS



Flesch-Kincaid Grade Level: 0 – 3: Kindergarten / Elementary, 3 – 6: Elementary, 6 – 9: Middle School, 9 – 12: High School, 12 – 15: College, 15 – 18: Post-graduate; Gunning Fog Index: 6- Sixth grade, 7- Seventh grade, 8- Eighth grade, 9-High school freshman, 10-High school sophomore, 11-High school junior, 12- High school senior, 13-College freshman, 14-College sophomore, 15-College junior, 16-College senior, 17 and above-College graduate. PLS, plain language summaries

LIMITATIONS

- The study findings are based on one generative AI tool (ChatGPT) and compare PLS within one therapeutic area (Oncology) with a limited sample size.
- ChatGPT may have intercepted data from the published PLS while generating the response.
- Figures and graphics, a key feature of PLS were not compared for quality.

References

- Pushparajah DS, et al. *Ther Innov Regul Sci*. 2018;52:474-481.
- Dormer L, et al. *Res Involv Engagem*. 2022;8:23.
- Altmäe S, et al. *Reprod Biomed Online*. 2023 ;47:3-9.
- Martin K. *Medical Writing*. 2023;32:22-27.

Acknowledgements

Roshan Kawale and Sandeep Chavan (SIRO Clinpharm Pvt. Ltd., India) provided graphic design assistance for the development of this poster.

Disclosures

Sonali Satam and Madhavi Patil are employees of SIRO Clinpharm Pvt. Ltd, India, Shweta Pitre and Priya Ganpathy are employees of SIRO Clinpharm UK Limited, UK and Vatsal Shah is an employee of SIRO Clinpharm USA LLC, NJ, USA.



Scan the QR code for the full digital poster

The QR code is intended to provide scientific information for individual reference, and the information should not be altered or reproduced in any way.