

Evaluating the quality of artificial intelligence-produced plain language summaries



OPEN HEALTH

Adam Fishbein,¹ Anna Battershill,¹
Lauren McNally²

¹OPEN Health, Parsippany, NJ, USA;

²OPEN Health, Marlow, Buckinghamshire, UK

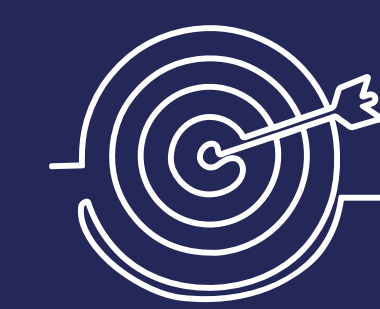
Objective

To assess the quality of plain language summaries (PLSs) produced by generative artificial intelligence (AI) versus medical writers (MWs)

Take-home messages



Our research suggests that, with appropriate prompting, **generative AI tools can produce high-quality PLSs** using abstracts from clinical articles



Although evaluators generally considered AI- and MW-produced PLSs to be of similar quality, **AI-produced PLSs were rated as significantly more accurate**

This may be because accuracy was judged in relation to each original abstract, and some MW-produced PLSs contained information that was not included in the original abstracts

Methods

- 1** We searched PubMed on September 25, 2023, for open access clinical articles that were published in English from 2020 onward, included a PLS, and disclosed MW support; 71 of the 78 articles that met these criteria were published by Future Science Group
- 2** We randomly selected 8 of the articles from Future Science Group journals and prompted Bing Chat Enterprise (which uses GPT-4) to produce a PLS for each article based on their abstracts
 - The prompt was developed by following prompt engineering guidance¹⁻³; it included key elements from the Future Science Group PLS guidelines
- 3** The existing MW-produced PLSs were extracted from the 8 articles to enable comparison between MW- and AI-produced PLSs
- 4** We developed evaluation forms to assess 5 aspects of each PLS (ie, domains): completeness, accuracy, understandability for a lay audience, clarity of language, and overall quality
 - Each domain was measured using a 5-point Likert scale, with higher scores indicating better evaluations; total scores (for all 5 domains) could range from 5–25
- 5** Forty OPEN Health staff completed subjective evaluations of 4 PLSs each
 - They were not informed that MW-produced PLSs were included in the research
 - Overall, 160 subjective evaluations were conducted: each of the 40 reviewers (20 scientific and 20 nonscientific background) evaluated 2 MW-produced and 2 AI-produced PLSs; each MW-produced PLS (n=8) and AI-produced PLS (n=8) was evaluated by 10 different reviewers
- 6** Quantitative measures (including readability scores) were recorded; inferential statistics (analyses of variance [ANOVA], *t* tests) were calculated using Graphpad Prism 9.5.1

- Mean text statistics and readability scores for MW- and AI-produced PLSs were similar (Table 1)

Table 1. Text statistics and readability scores

Measures, mean (standard deviation)	MW-produced PLSs (n=8)	AI-produced PLSs (n=8)	P-value ^a
Word count	287 (26.7)	252 (36.5)	0.119
Complex words %	15.8 (2.8)	15.8 (4.1)	0.959
Passive voice sentences %	23.3 (14.1)	16.1 (5.0)	0.108
Flesch-Kincaid reading ease score ^b	49.4 (8.0)	52.1 (10.3)	0.485
Flesch-Kincaid grade level score ^c	11.0 (1.7)	10.1 (1.5)	0.210

Most measures were calculated using <https://www.analyzemymwriting.com/>; <https://www.webfx.com/tools/read-able> was used to calculate Flesch-Kincaid reading ease. ^aAssessed using paired *t* tests. ^bBased on a ranking scale of 0–100; the higher the score, the easier to read. ^cAmerican school grade needed to comprehend the material; the lower the score, the easier to read.

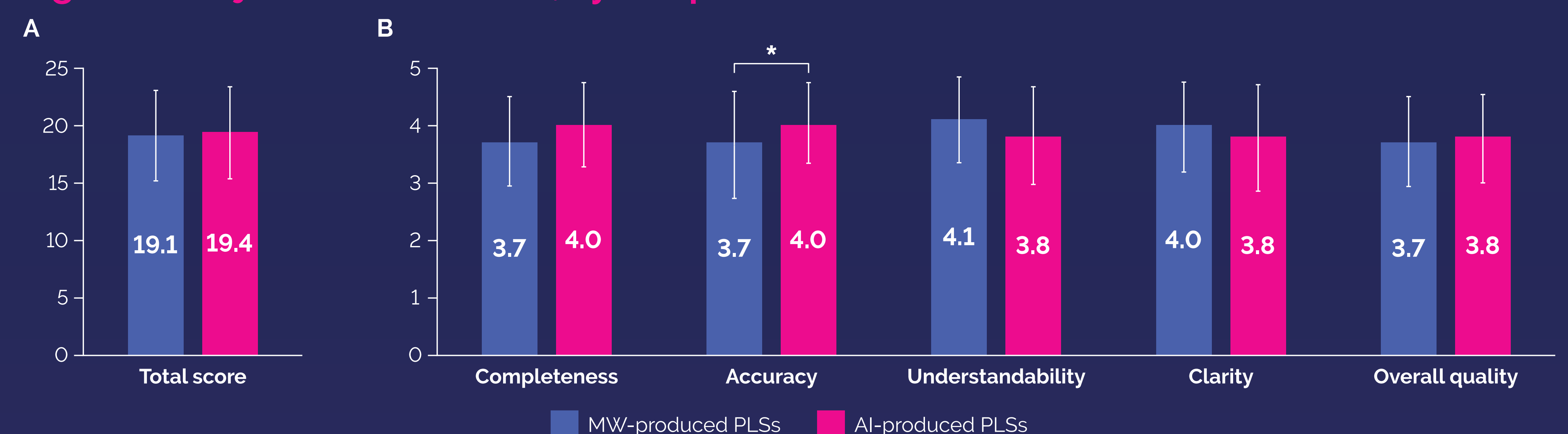
AI, artificial intelligence; MW, medical writer; PLS, plain language summary.

Acknowledgments: PLSs that were evaluated as part of this research were generated using Bing Chat Enterprise (Microsoft). The authors thank their OPEN Health colleagues Kaydee Wicks and Kyle Gies for creative and design assistance and Jill Sebra and Teri O'Neill for editorial support.

References: 1. OpenAI developer platform. Prompt engineering. <https://platform.openai.com/docs/guides/prompt-engineering>. 2. Best practices for prompt engineering with the OpenAI API. <https://help.openai.com/en/articles/6654000-best-practices-for-prompt-engineering-with-the-openai-api>. 3. Master prompt engineering: demystifying prompting through a structured approach. Prompt Engineering and AI Institute. <https://promptengineering.org/master-prompt-engineering-demystifying-prompting-through-a-structured-approach>.

- Mean total scores obtained from subjective evaluations were similar for MW- and AI-produced PLSs (Fig. 1A), although AI-produced PLSs were rated significantly higher in accuracy (F[1,76]=4.496; *P*=0.037) (Fig. 1B)

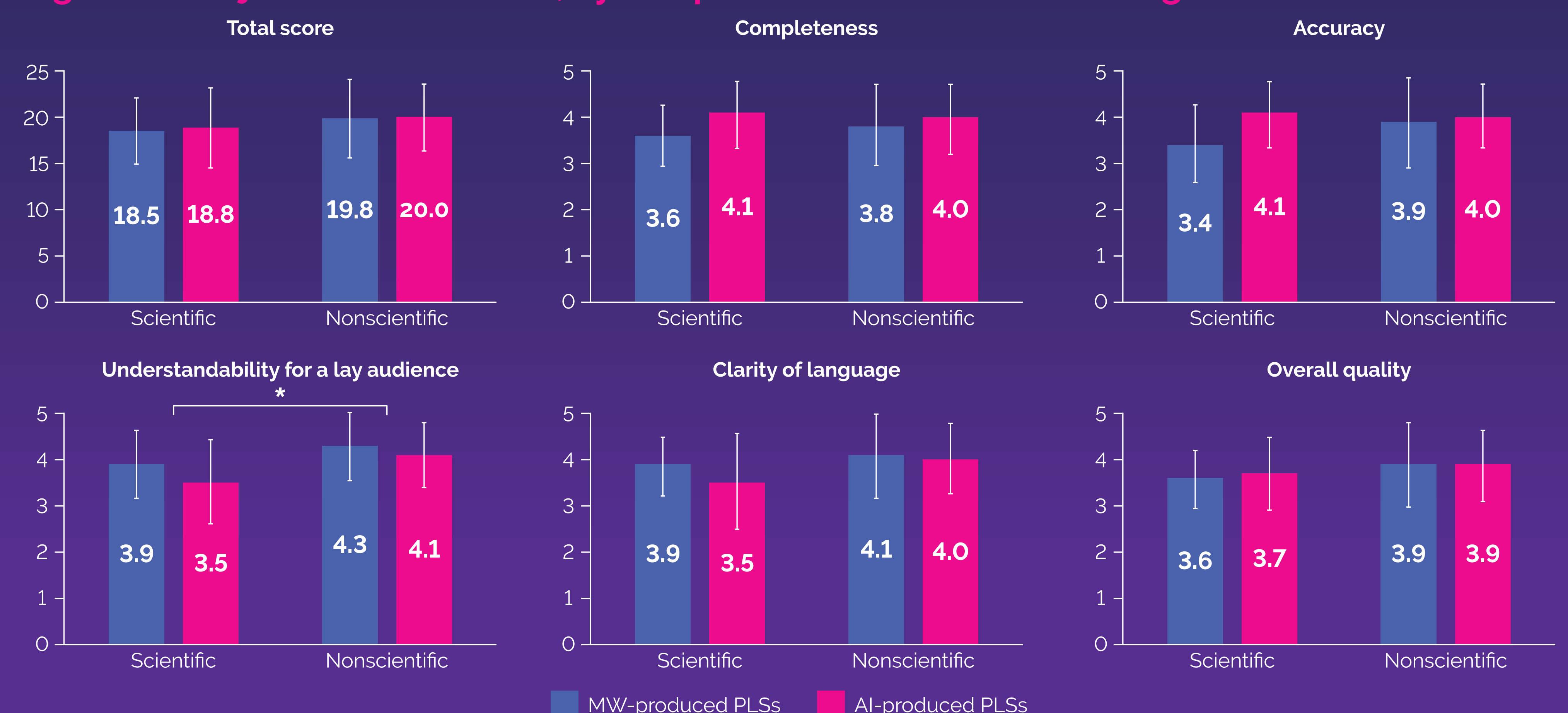
Figure 1. Subjective evaluations, by PLS producer



**P*<0.05 (assessed using 2-way ANOVA [two factors: PLS producer and reviewer background]). Data are mean (standard deviation). AI, artificial intelligence; MW, medical writer; PLS, plain language summary.

- No significant interactions between PLS producer (MW vs AI) and reviewer background (scientific vs nonscientific) were observed for any of the subjective evaluation domains (Fig. 2); however, in terms of understandability for a lay audience, reviewers with a nonscientific background rated PLSs significantly higher than reviewers with a scientific background (F[1,76]=7.525; *P*=0.008)

Figure 2. Subjective evaluations, by PLS producer and reviewer background



**P*<0.01 (assessed using 2-way ANOVA). Data are mean (standard deviation). Reviewers with scientific versus nonscientific backgrounds were evenly split (n=20 each), and each PLS was evaluated by an equal number of scientific and nonscientific reviewers. AI, artificial intelligence; MW, medical writer; PLS, plain language summary.

Author statement: All authors attest to participating in the research process or development of the concept discussed, were actively involved in writing the abstract and poster, and approved the submitted/presented versions.

Acknowledgments: PLSs that were evaluated as part of this research were generated using Bing Chat Enterprise (Microsoft). The authors thank their OPEN Health colleagues Kaydee Wicks and Kyle Gies for creative and design assistance and Jill Sebra and Teri O'Neill for editorial support.

References: 1. OpenAI developer platform. Prompt engineering. <https://platform.openai.com/docs/guides/prompt-engineering>. 2. Best practices for prompt engineering with the OpenAI API. <https://help.openai.com/en/articles/6654000-best-practices-for-prompt-engineering-with-the-openai-api>. 3. Master prompt engineering: demystifying prompting through a structured approach. Prompt Engineering and AI Institute. <https://promptengineering.org/master-prompt-engineering-demystifying-prompting-through-a-structured-approach>.

Scan me to download supplementary materials (including the AI prompt and the PLSs evaluated)

